

Ai confini del vivente

Ciclo monografico sulla biologia, l'evoluzione e il fondamento organico dell'intelligenza

Introduzione generale al ciclo

Il secondo programma: dal calcolabile al vivente

Il ciclo *Ai confini del calcolabile* aveva costruito la mappa formale dell'intelligenza artificiale tracciando i confini di ciò che qualsiasi sistema computazionale può fare. La pars destruens aveva dimostrato, con la definitività dei teoremi di impossibilità, dove si trovano i muri insuperabili: l'incompletezza di Gödel, l'indcidibilità di Turing, l'impossibilità di Arrow, l'explanatory gap di Chalmers, il limite della gerarchia causale di Pearl. La pars costruens aveva mostrato come abitare intelligentemente il territorio delimitato da quei muri: la razionalità ecologica, la causalità interventistica, la cooperazione evolutiva, l'architettura dei valori. Il terzo movimento aveva infine esplorato le direzioni aperte: il territorio tra Gödel e Bayes, la teoria funzionale della decisione, i giochi evolutivi cognitivi, l'irriducibilità del normativo.

Il ciclo si era chiuso con la consapevolezza di un programma incompleto in un senso preciso: le ventiquattro monografie sulla calcolabilità avevano trattato l'intelligenza come oggetto formale — come sistema di assiomi, funzioni, algoritmi, distribuzioni di probabilità — senza mai affrontare la domanda di fondo che ogni tentativo di costruire intelligenza artificiale presuppone senza rispondere: che cos'è l'intelligenza biologica che l'intelligenza artificiale cerca di emulare, approssimare o superare? Da dove viene? Come funziona nei dettagli del suo substrato materiale? Quali proprietà del substrato biologico sono necessarie, e quali accidentali?

Ai confini del vivente è il tentativo di rispondere a questa domanda con la stessa precisione formale che il ciclo precedente aveva applicato alla questione computazionale. Il programma è radicalmente diverso: non assiomi e algoritmi ma neuroni e sinapsi, non teoremi di impossibilità

ma limiti di Hayflick, non ottimizzazione stocastica ma selezione naturale, non macchine di Turing ma cellule, tessuti, corpi. Eppure il metodo è identico: costruire il modello formale più preciso disponibile, inferirne le proprietà con rigore matematico, confrontarle con l'evidenza sperimentale, e tracciare i limiti di ciò che il modello può spiegare e di ciò che strutturalmente eccede ogni spiegazione formale.

Il titolo *Ai confini del vivente* ha una duplice valenza. Da un lato indica il confine esterno del programma: la neurobiologia, l'evoluzione e la biologia molecolare portate ai loro limiti descrittivi — il punto in cui la descrizione meccanicistica del sistema biologico raggiunge i propri bordi e si imbatte nelle domande che non sa risolvere con i propri strumenti. Dall'altro indica il confine interno del vivente: la struttura biologica stessa come sistema che opera al confine tra l'ordine e il disordine, tra la determinazione e la stocasticità, tra la replica e l'innovazione, tra il codice e la sua realizzazione materiale. Il vivente è costitutivamente un sistema di confine — e studiarlo significa studiare la struttura formale del confine stesso.

La struttura del ciclo in tre movimenti

Il ciclo è organizzato in tre movimenti che rispecchiano, con una logica propria, la struttura del ciclo precedente.

Il primo movimento (Monografie I–VIII): la geometria del biologico. Il primo movimento costruisce i fondamenti matematici della biologia dell'intelligenza. La Monografia I apre con la geometria dell'apprendimento — la dimensione VC, il teorema di sovrapposizione, il binding problem — e stabilisce che la cognizione neurale non è una funzione generica ma un'operazione su strutture geometriche di alta dimensionalità con proprietà formali precise. La Monografia II introduce il motore evolutivo: l'equazione di Price come teorema di invarianza dell'evoluzione per selezione, i modelli NK di Kauffman come paesaggi adattativi, il teorema del No Free Lunch come vincolo universale sull'ottimizzazione. Le Monografie III e IV affrontano la percezione come inferenza bayesiana gerarchica e la struttura a tre livelli di Marr — la distinzione tra computazione, algoritmo e implementazione che è il principio epistemologico unificante dell'intero ciclo. Le Monografie V e VI costruiscono la teoria dello sviluppo — la genetica dello sviluppo, le reti regolatorie dei geni, la biologia del cancro come sistema evolutivo intracellulare. La Monografia VII introduce la neurobiologia della cognizione spaziale — le cellule luogo, le cellule griglia, la navigazione come computazione geometrica — come caso paradigmatico del rapporto tra substrato materiale e funzione cognitiva. La Monografia VIII chiude il primo movimento con la struttura del

sistema nervoso autonomo e l'interocezione, anticipando il fondamento corporeo dell'emozione che il terzo movimento svilupperà formalmente.

Il secondo movimento (Monografie IX–XVI): il fondamento organico del valore. Il secondo movimento sposta il fuoco dalla percezione alla valutazione, dall'elaborazione dell'informazione all'orientamento motivazionale del comportamento. La Monografia IX introduce la neurobiologia del sistema dopaminergico — il temporal difference learning come algoritmo di apprendimento del valore, il ruolo della dopamina come segnale dell'errore di predizione del rinforzo. La Monografia X affronta la struttura dei gangli della base — il circuito cortico-striato-talamo-corticale, la selezione dell'azione come ottimizzazione in circuiti inibitorio-eccitatori. La Monografia XI introduce la neurobiologia della cura e dell'attaccamento — i neuropeptidi (ossitocina, vasopressina), le basi biologiche del legame, la connessione tra neurobiologia del sociale e sistemi di valore. Le Monografie XII e XIII affrontano la neurobiologia del linguaggio e del pensiero simbolico — la struttura anatomica delle aree di Broca e Wernicke, la connessione tra sintassi e sequenza motoria, l'ipotesi neuromotoria del linguaggio. La Monografia XIV costruisce la neurobiologia dell'etica e della moralità — la doppia elaborazione di Greene, il ruolo del cingolato anteriore e della corteccia prefrontale ventromediale nelle decisioni morali. La Monografia XV introduce la biologia della coscienza a livello implementazionale — il ruolo del talamo come struttura di integrazione talamocorticale, le onde corticali lente, il sistema di attivazione reticolare. La Monografia XVI chiude il secondo movimento con la neurobiologia dello sviluppo cognitivo e della plasticità sinaptica nella finestra critica — come l'esperienza modella strutturalmente i circuiti durante i periodi sensibili dello sviluppo.

Il terzo movimento (Monografie XVII–XXIV): ai confini della mente. Il terzo movimento affronta le questioni più profonde e formalmente più impegnative: la struttura della mente bayesiana e le sue approssimazioni neurali, la neurobiologia della coscienza e il confronto tra le grandi teorie, la semantica evolutiva dall'origine della comunicazione molecolare al linguaggio umano ricorsivo, la struttura formale della memoria con i meccanismi di consolidazione e riconsolidazione, la fisica dell'emozione come termodinamica degli stati affettivi, i giochi evolutivi della neurobiologia nella corsa agli armamenti immunitari e nella coevoluzione del cervello sociale, i sistemi biologici di error correction dall'UPR alla biologia dei telomeri, e infine la struttura fenomenologica del corpo vissuto come confine irriducibile di ogni programma riduzionistico.

Le connessioni interne tra i tre movimenti

I tre movimenti non sono indipendenti: formano un sistema con connessioni strutturali che questa introduzione vuole rendere esplicite.

La connessione tra il primo e il secondo movimento è quella tra *computazione* e *valutazione*. Il primo movimento costruisce la macchina che elabora l'informazione; il secondo costruisce il sistema che assegna valore ai risultati di quell'elaborazione. Questa connessione non è contingente — non è che il sistema valutativo si aggiunge alla macchina computazionale come modulo separato. Il secondo movimento mostra che la valutazione è intrecciata nella struttura computazionale a ogni livello: il sistema dopaminergico modifica i pesi sinaptici attraverso gli stessi meccanismi di plasticità che il primo movimento ha analizzato come apprendimento hebbiano; i gangli della base selezionano le sequenze di azione attraverso un circuito inibitorio-eccitatorio che implementa la stessa struttura dell'ottimizzazione per discesa del gradiente. La valutazione non è sopra la computazione: è nella computazione.

La connessione tra il secondo e il terzo movimento è quella tra *valore* e *significato*. Il secondo movimento costruisce il sistema neurologico del valore come orientamento motivazionale del comportamento; il terzo movimento mostra come questo sistema si articola in strutture di significato — la mente bayesiana come inferenza guidata dal valore atteso, il linguaggio come sistema di comunicazione del valore atteso agli altri, la coscienza come esperienza integrata che include la dimensione valutativa come componente irriducibile, le emozioni come modulatori della precisione nella computazione inferenziale. Il significato non è separabile dal valore: nel sistema biologico, comprendere qualcosa significa computarne la rilevanza per i propri scopi — una struttura che unifica la semantica e la pragmatica in modo che le teorie formali del significato faticano a catturare.

La connessione tra il primo e il terzo movimento è quella tra *geometria* e *fenomenologia*. Il primo movimento ha costruito la geometria dell'apprendimento — gli spazi ad alta dimensionalità, le varietà differenziali, le strutture topologiche della rappresentazione neurale. Il terzo movimento mostra che questa geometria ha un'altra faccia: la struttura fenomenologica del corpo vissuto — lo schema corporeo come geometria incarnata, i neuroni specchio come substrato della risonanza geometrica con i corpi altrui, il chiasmo di Merleau-Ponty come struttura della reversibilità tra il toccante e il toccato. La geometria della cognizione neurale e la geometria fenomenologica del corpo vissuto sono due descrizioni dello stesso oggetto da due lati opposti del confine che separa la terza persona dalla prima.

Il tema unificante: la biologia come implementazione dell'inferenza

Se il tema unificante del ciclo precedente era *l'irriducibilità del normativo* — la scoperta che ogni dominio del ciclo contiene un livello normativo non riducibile né alla descrizione né alla formalizzazione — il tema unificante del presente ciclo è quello che potremmo chiamare *l'inferenza incarnata*: la scoperta che il sistema biologico, a ogni livello di organizzazione dalla singola cellula all'organismo nella sua integralità sociale, implementa strutture di inferenza — aggiornamento di credenze alla luce di evidenza, modulazione dell'incertezza, ottimizzazione vincolata — con una profondità e una pervasività che trascende ogni singolo sistema o funzione.

La cellula inferisce la concentrazione di nutrienti dalla distribuzione delle molecole-segnalet. L'embrione inferisce la propria posizione nello spazio morfogenetico dal gradiente dei morfogeni. Il sistema immunitario inferisce la presenza di patogeni dalla geometria molecolare degli epitopi. Il sistema nervoso inferisce lo stato del mondo dalla struttura degli errori di previsione nella gerarchia corticale. Il soggetto cosciente inferisce il significato del discorso altrui dalla risonanza del proprio sistema motorio con i movimenti del parlante. A ogni livello, l'inferenza biologica è più ricca, più flessibile, più robusta ai disturbi e più economica in termini di risorse computazionali di qualsiasi implementazione artificiale.

Questa pervasività dell'inferenza biologica non è un accidente evolutivo — è il risultato di tre miliardi di anni di selezione per sistemi che aggiornano le proprie credenze in modo adattativo in ambienti stocastici. L'evoluzione ha selezionato i sistemi biologici per essere macchine inferenziali ottimali nel senso variazionale: sistemi che minimizzano l'energia libera variazionale della propria rappresentazione del mondo, che massimizzano la precisione delle proprie previsioni sensoriali, che bilanciano la semplicità del modello con la sua accuratezza predittiva. Il principio di energia libera di Friston — il tema unificante della Monografia XVII — non è soltanto una teoria della percezione: è una teoria di come i sistemi biologici siano costruiti evolutivamente per inferire.

Le differenze strutturali con il ciclo precedente

Il ciclo *Ai confini del vivente* differisce dal precedente in tre modi strutturali che è utile rendere espliciti prima di procedere alla lettura.

Il primo è la *relazione con l'evidenza sperimentale*. Il ciclo precedente operava prevalentemente nel dominio dei sistemi formali — le prove erano dimostrazioni matematiche, i risultati erano teoremi. Il presente ciclo opera in un dominio in cui la matematica e la biologia sperimentale si intersecano continuamente: i modelli formali derivano le proprie strutture dall'evidenza sperimentale, e le predizioni formali vengono confrontate con misurazioni empiriche precise. Ogni sezione costruisce il modello e poi lo ancora ai dati: le costanti di smorzamento degli stati affettivi di Kuppens et al., i profili laminari dell'attivazione corticale di Bastos et al., il replay ippocampale di Wilson e McNaughton, la cinesi dei tre sensori UPR nelle cellule in stress. Il formalismo senza il dato è vuoto; il dato senza il formalismo è cieco.

Il secondo è la *struttura dei limiti*. Il ciclo precedente tracciava limiti assoluti — teoremi di impossibilità, confini insuperabili. Il presente ciclo traccia limiti di tipo diverso: non impossibilità logiche ma vincoli termodinamici, non indecidibilità computazionale ma limiti di Hayflick, non teorema di Rice ma invecchiamento come accumulo irriducibile di errori irrilevanti. I limiti biologici sono più sfumati di quelli computazionali: non sono barriere verticali ma pendii — confini che si raggiungono gradualmente, non improvvisamente. Questa differenza non li rende meno fondamentali: il limite di Hayflick non è meno strutturale del teorema di incompletezza di Gödel. Ma la struttura del confine è diversa, e richiede strumenti matematici diversi per essere descritta con precisione.

Il terzo è la *presenza del confine fenomenologico*. Il ciclo precedente aveva affrontato il problema difficile della coscienza come impossibilità epistemica — il divario esplicativo di Chalmers come confine tra la descrizione funzionale e l'esperienza soggettiva. Il presente ciclo affronta questo confine non come impossibilità astratta ma come struttura concreta del corpo vissuto: il chiasmo di Merleau-Ponty, i neuroni specchio come substrato dell'*embodied simulation*, l'intercorporeità come costituzione sociale dello schema corporeo. La fenomenologia qui non è un'aggiunta filosofica esterna alla biologia: è il modo in cui il sistema biologico si manifesta dalla propria prospettiva interna, e questa prospettiva interna è tanto parte del fenomeno quanto la sua descrizione neurofisiologica dall'esterno.

Una nota sul metodo

La struttura di ciascuna monografia segue un metodo invariante che è utile descrivere una volta per tutte. Ogni monografia parte da una domanda empiricamente motivata — non da un assioma arbitrario ma da un problema che la biologia e la neuroscienza hanno identificato come fondamentale. La domanda viene poi formalizzata: le entità biologiche rilevanti vengono rappresentate come oggetti matematici (vettori, funzioni, distribuzioni, grafi, equazioni differenziali), e le relazioni tra di esse vengono espresse come relazioni matematiche. Dal modello formale vengono derivate proprietà — teoremi, predizioni quantitative, strutture di equilibrio. Le proprietà derivate vengono confrontate con l'evidenza sperimentale disponibile. Infine, i limiti del modello vengono tracciati: che cosa spiega, che cosa presuppone senza spiegare, e che cosa strutturalmente eccede la sua capacità descrittiva.

Questo metodo — domanda $\rightarrow$ formalizzazione $\rightarrow$ derivazione $\rightarrow$ confronto $\rightarrow$ limite — è lo stesso che ha guidato il ciclo precedente. La differenza è che nel ciclo sulla calcolabilità la formalizzazione era spesso data a priori (sistemi formali, macchine di Turing, algoritmi), mentre nel presente ciclo la formalizzazione è sempre un'operazione costruttiva che richiede scelte di modellazione non triviali. Scegliere di rappresentare gli stati emotivi come un sistema dinamico nel piano valenza-arousal, o di rappresentare l'architettura dell'amigdala come un circuito $BLA \rightarrow ICN \rightarrow CeA$ con soglie e proiezioni agli effettori, sono scelte teoriche che determinano le predizioni del modello e i suoi limiti. Queste scelte vengono sempre rese esplicite e motivate dall'evidenza sperimentale che le sostiene.

Sommario delle Monografie I–XVI

Le sedici monografie dei primi due movimenti vengono richiamate nel terzo movimento con riferimenti espliciti. Il sommario seguente offre i risultati tecnici centrali di ciascuna nella forma necessaria per comprendere i richiami.

Monografia I: La geometria dell'apprendimento neurale. Teorema di Minsky-Papert: una singola soglia lineare non separa la funzione XOR; dimensione n richiede $n+1$ strati per separare funzioni di parità. Dimensione VC: la classe delle funzioni lineari in $\mathbb{R}^n$ ha $VC\text{-dim} = n+1$. Teorema di superposizione di Kolmogorov: qualsiasi funzione continua $f: [0,1]^n \rightarrow \mathbb{R}$ può essere espressa come $f(x_1, \dots, x_n) = \sum_{q=1}^{2n+1} g_q(\sum_{p=1}^n h_{q,p}(x_p))$ con h e g continue. Il binding problem: la sincronia oscillatoria a 40 Hz (oscillazioni gamma) come soluzione temporale al problema della separazione dei feature set distribuiti. I

campi recettivi come kernel di Gabor con relazione di incertezza $\Delta x \cdot \Delta \omega \geq 1/2$.

Monografia II: Il motore evolutivo. Equazione di Price: $\Delta \bar{z} = \text{Cov}(w_i, z_i)/\bar{w} + E(w_i \Delta z_i)/\bar{w}$ — il cambiamento nella media fenotipica è la somma della selezione (covarianza fitness-fenotipo) e della trasmissione (aspettativa del cambiamento pesato per la fitness). Modello NK: un sistema con N elementi ciascuno con K interazioni ha 2^N stati con un paesaggio adattativo la cui rugosità cresce con K; per $K=0$ esiste un unico massimo globale, per $K=N-1$ il paesaggio è massimamente caotico. Teorema del No Free Lunch (Wolpert e Macready, 1997): per qualsiasi coppia di algoritmi di ottimizzazione A_1, A_2 , la somma delle performance su tutti i possibili problemi di ottimizzazione è identica — non esiste un algoritmo superiore su tutti i possibili problemi.

Monografia III: La percezione come inferenza bayesiana. Aggiornamento bayesiano gerarchico: $P(z_l | x) \propto P(x | z_l) \cdot \prod_l P(z_l | z_{l+1}) \cdot P(z_L)$. Predictive coding (Friston, 2005): errore di previsione $\epsilon_l = (x_l - f_l(\mu_{l+1})) / \Sigma_l$. Aggiornamento della rappresentazione: $d\mu_l/dt = -\kappa(\epsilon_l - f'_l(\mu_{l+1})^T \cdot \epsilon_{l-1})$. ELBO: $\log P(x) \geq E_q[\log P(x,z)] - E_q[\log q(z)] = -F(q)$. Integrazione multisensoriale ottimale: $\hat{s} = (\sigma_v^2 \cdot s_v + \sigma_t^2 \cdot s_t) / (\sigma_v^2 + \sigma_t^2)$ con $\sigma_{\text{comb}}^2 = \sigma_v^2 \sigma_t^2 / (\sigma_v^2 + \sigma_t^2)$ (Ernst e Banks, 2002).

Monografia IV: I livelli di Marr. Il principio dei tre livelli: computazione (che problema viene risolto?), algoritmo (come viene risolto?), implementazione (su quale substrato fisico?). Indipendenza dei livelli: un'analisi completa al livello dell'implementazione non determina la descrizione al livello computazionale, e viceversa. Il programma riduzionista come strategia di ricerca, non come tesi ontologica: la riduzione al livello dell'implementazione è sempre incompleta rispetto al livello computazionale — il quale rimane sempre aperto al divario esplicativo di Chalmers.

Monografia V: Le reti regolatorie dei geni. Rete di Kauffman NK come Boolean Network: la stabilità richiede $K \leq 2$ (corrispondente alle reti di cablaggio sparso delle GRN biologiche). Teorema di stabilità: per $K = 2$, la lunghezza media degli attrattori è $\sqrt{N}$ — compatibile con il numero di tipi cellulari osservati nei mammiferi ($\sim 256 \approx \sqrt{N}$ per $N \approx 10^5$). Modello di Hill per la trascrizione: $f(x) = x^n / (K^n + x^n)$ con la cooperatività come meccanismo di bistabilità. L'immunoediting tumorale: Eliminazione $\rightarrow$ Equilibrio $\rightarrow$ Evasione come sistema di tre fasi con dinamiche formalmente analoghe al modello preda-predatore Lotka-Volterra.

Monografia VI: La neurobiologia del sistema dopaminergico.

TD-learning: $V(s_t) \leftarrow V(s_t) + \alpha \cdot \delta_t$ dove $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ è l'errore di predizione temporale. Il segnale della dopamina come δ : la scarica fasica dei neuroni dopaminergici del VTA corrisponde a $\delta > 0$, la soppressione corrisponde a $\delta < 0$ (Schultz, 1997). Actor-Critic con eligibility traces: $\Delta w_{actor} = \alpha_a \cdot \delta \cdot e_{actor}$, $\Delta V(s) = \alpha_c \cdot \delta \cdot e_{critic}$.

Monografia VII: Cellule luogo, cellule griglia e la navigazione come computazione geometrica.

Le cellule luogo di O'Keefe (1971): scariche selettive per posizioni nell'ambiente. Le cellule griglia di Moser (2005): pattern esagonali di scarica che formano reticoli con periodicità multipla. La rappresentazione vettoriale della navigazione: il percorso integrale $\Delta r(t) = \int v(t') dt'$ implementato dal circuito entorinale-ippocampale. Il modello di grid cells come oscillatori accoppiati con interferenza di fase.

Monografia VIII: L'interocezione e il fondamento corporeo della mente.

La corteccia insulare anteriore come sito della codifica interolettiva predittiva. Il ciclo corpo-cervello-corpo: gli stati somatici come implementazione dei marcatori somatici di Damasio. Il principio della predictive allostasis: il sistema interolettivo non reagisce agli stati corporei ma li prevede e li regola anticipatoriamente (Sterling, 2012).

Monografia IX: I gangli della base come sistema di selezione dell'azione.

Il circuito cortico-striato-talamo-corticale: Input corticale $\rightarrow$ Striato $\rightarrow$ GPi/SNpr (via diretta: disinibizione; via indiretta: inibizione aggiuntiva tramite GPe-STN) $\rightarrow$ Talamo $\rightarrow$ Corteccia. La competizione tra percorsi diretti e indiretti come meccanismo di selezione dell'azione. La regola di aggiornamento sinaptico cortico-striale: $\Delta w = \eta \cdot \delta \cdot e$ dove δ è il segnale dopaminergico e e è la traccia di eligibilità.

Monografia X: La neurobiologia del legame e dell'attaccamento.

Il ruolo dell'ossitocina e della vasopressina nel legame paio (Insel, 1997). La distribuzione dei recettori V1aR nel nucleus accumbens come correlato del comportamento di coppia monogama nelle vole di prateria. Il modello formale del legame come abitudine dopaminergica con rinforzo sociale: $\Delta w_{\{legame\}} = \eta_{social} \cdot \delta \cdot e$.

Monografia XI: La neurobiologia del linguaggio.

Il doppio percorso del linguaggio (Hickok e Poeppel, 2004): percorso dorsale (suono $\rightarrow$ articolazione, via parieto-frontale) e percorso ventrale (suono $\rightarrow$ significato, via temporale inferiore). La connessione con la neurobiologia motoria: la struttura delle aree di Broca e Wernicke come estensione delle aree del gesto manuale e del riconoscimento biologico del movimento.

Monografia XII: La neurobiologia dell'etica e della moralità. Il modello di doppia elaborazione di Greene (2001): risposta emotiva rapida (coinvolge amigdala, vmPFC, cingolato posteriore) versus deliberazione utilitaristica (coinvolge dlPFC, cingolato anteriore). La risoluzione del conflitto tra i due sistemi come soppressione della risposta emotiva da parte della dlPFC attraverso inibizione discendente.

Monografia XIII: La struttura del cervello sociale. Connessioni corticali della mentalizzazione: corteccia prefrontale mediale (mPFC), giunzione temporoparietale (TPJ), corteccia temporale posteriore. La rete del default mode come substrato della simulazione mentale e della prospettiva in prima persona.

Monografia XIV: La plasticità sinaptica e le finestre critiche. La finestra critica dello sviluppo come periodo di plasticità potenziata mediata da fattori molecolari (BDNF, PSA-NCAM, PNN). La chiusura della finestra critica come consolidazione dell'architettura sinaptica attraverso la formazione delle perineuronal nets. Il modello formale dell'interazione esperienza-genetica come epigenesi probabilistica: $P(\text{connessione} \mid \text{esperienza, genetica})$.

Monografia XV: Il talamo come hub integrativo della coscienza. Il modello talamocorticale della coscienza: il talamo come relay selettivo che amplifica o sopprime l'attività corticale attraverso l'oscillazione tra stati up (depolarizzato, eccitabile) e down (iperpolarizzato, silente). I nuclei talamici specifici (relay sensoriale) vs. non specifici (modulazione diffusa). Il ruolo del nucleo reticolare talamico come gate inibitorio del flusso di informazioni talamocorticali.

Monografia XVI: La neurobiologia dello sviluppo cognitivo. L'arborizzazione dendritica guidata dall'attività: la regola di Hebbian sprouting e pruning. La mielinizzazione come ottimizzazione della velocità di conduzione con vincoli energetici. La traiettoria di maturazione corticale nei primati: la corteccia prefrontale è l'ultima area a maturare (fino ai 25 anni negli esseri umani), con implicazioni per la cognizione esecutiva e il controllo degli impulsi durante l'adolescenza.

Sommario delle Monografie XVII–XXIV (Terzo Movimento)

Monografia XVII: La mente bayesiana. Il programma della mente bayesiana come inferenza variazionale su modelli generativi gerarchici. $ELBO = E_q[\log P(x,z)] - E_q[\log q(z)]$; minimizzazione dell'energia libera variazionale $F = -ELBO$. La stratificazione laminare del

predictive coding: strati II/III $\rightarrow$ segnali di errore feedforward in banda gamma; strati V/VI $\rightarrow$ previsioni feedback in banda beta (Bastos et al. 2012, 2015). La psicopatologia come disregolazione della precisione: psicosi ($\Pi_s^{\text{psicosi}} \gg \Pi_s^{\text{normale}}$, iperprecisione degli errori sensoriali); ansia (circolo autorinforzante di iperprecisione enterocettiva); depressione (ipoprecisione dei segnali di ricompensa, γ_v ridotto). AIXItl come versione computabile di AIXI: $M_{\{t\}}(x_1 \dots x_n) = \sum_{p: U(p,t)=x, |p| \leq l} 2^{-|p|}$; teorema di convergenza: AIXItl $\rightarrow$ AIXI per $l, t \rightarrow \infty$.

Monografia XVIII: La neurobiologia della coscienza. IIT 3.0: causa purview $\pi^c(s \rightarrow M) = p(M_{\{t-1\}} | do(s_t = s_c))$; effetto purview $\pi^e(s \rightarrow M) = p(M_{\{t+1\}} | do(s_t = s_c))$; phi del meccanismo $\varphi(s) = EMD(\pi^c\{ce\}(s \rightarrow M), \pi^e\{ce\}\{MIP\}(s \rightarrow M))$; *struttura concettuale CES(M)* = $\{(s, \varphi(s), \pi^c\{ce\}(s \rightarrow M)) : s \subseteq M, \varphi(s) > 0\}$; $\Phi(M) = \min\{P\} D(CES(M), CES_P(M))$. Global Workspace Theory: ignizione non lineare come transizione di fase in un sistema dinamico $dr/dt = -r + F(W_{ff} \cdot s + W_{lat} \cdot r)$; onda P3b come firma dell'ignizione. HOT di Rosenthal vs HOP di Lycan: HOT richiede meta-rappresentazione proposizionale; HOP richiede soltanto monitoring mechanism procedurale. Adversarial collaboration COGITATE (Melloni et al., 2023): localizzazione posteriore degli NCC favorita da sEEG, MEG e fMRI; attivazione prefrontale correlata con il report piuttosto che con l'esperienza fenomenica. $PCI = C_{\{LZ\}}(risposta_{\{TMS\}}) / \log_2(n_t \cdot n_e) > 0.31$ in tutti gli stati coscienti.

Monografia XIX: La semantica evolutiva. Grammatiche CF vs regolari: le grammatiche context-free (Chomsky tipo 2) generano le strutture center-embedded del linguaggio umano; le grammatiche regolari (tipo 3) sono insufficienti. Studio Fitch-Hauser 2004: gli scimpanzé apprendono pattern regolari $A^n B^n$ ma non strutture CF con embedding. FLN (Hauser-Chomsky-Fitch, 2002): la ricorsività come unico tratto specificamente umano nella facoltà del linguaggio in senso stretto. FOXP2: le due sostituzioni umane (T303N e N325S) rispetto allo scimpanzé mostrano segnali di selezione positiva rapida; esperimento Enard 2009 (topi $FOXP2^{\text{humanized}}$): vocalizzazione più varia, plasticità striatale aumentata, apprendimento procedurale più rapido. Semantica di Montague: tipi $\langle e, t \rangle$ (proprietà), $\langle \langle e, t \rangle, t \rangle$ (quantificatori); regola di applicazione funzionale $[[NP \ VP]] = [[NP]]([VP])$; semantica intensionale con mondi possibili $s: \langle s, \langle e, t \rangle \rangle$ per proprietà intensionali. Semantica degli eventi di Davidson: $\exists e [azione(e, s) \wedge strumento(e, i) \wedge luogo(e, l)]$. Problema dell'iperintensionalità e semantica strutturata.

Monografia XX: La struttura formale della memoria. Regola di Hebb con normalizzazione di Oja: $\Delta w_{\{ij\}} = \eta \cdot x_i \cdot (x - x_i \cdot w_i)$. BCM theory: $dw/dt = \varphi(x_{\text{post}}, \theta_M) \cdot x_{\text{pre}}$ con $d\theta_M/dt = \langle x_{\text{post}} \rangle -$

$\theta_M)/\tau_\theta$. Synaptic tagging and capture (Frey-Morris 1997): $\Delta w_{\{L-LTP\}}(s) = [\text{tag}(s,t) > 0] \cdot [\text{PRP_conc}(t) > \theta_{\{\text{PRP}\}}]$; finestra di integrazione tagging-capture: 1-2 ore. Oscillazioni theta (4-8 Hz): fase encoding a 270° , fase retrieval a 90° — separazione di 180° per prevenire interferenza. Sharp-wave ripple: compressione del replay ippocampale di $\sim 6000\times$ (Wilson-McNaughton, 1994). Dialogue hypothesis: SO (0.5 Hz) $\rightarrow$ Fusi (12 Hz) $\rightarrow$ SWR (100-250 Hz). Reconsolidazione (Nader-Schafe-LeDoux 2000): de-consolidazione mediata da AMPAR endocitosi, calpain e MMP-9; riconsolidazione mediata da sintesi proteica via PKA-CREB. Propranololo come modulatore terapeutico della riconsolidazione emotiva nell'amigdala. Rete di Hopfield in CA3: capacità $M_{\max} \approx 0.138 \cdot N$; separazione dei pattern nel DG con sparsità $a \approx 0.02$.

Monografia XXI: La fisica dell'emozione. Somatic Marker Hypothesis: $V_{\text{scelta}} = \text{argmax}_{\{O \in \text{Candidati}\}} [V_e(O) + \lambda \cdot V_i(O)]$ con $\text{Candidati} = \{O : V_i(O) > \theta_s\}$. Component Process Model: Valenza $\approx \beta_1 \cdot \text{congruenzaScopi} + \beta_2 \cdot \text{gradevole}$; Arousal $\approx \gamma_1 \cdot \text{novità} + \gamma_2 \cdot \text{potenzialeCoping}^{-1}$. Potenziale affettivo $U(v,a)$: $dv/dt = -\gamma_v(v-v_{\text{SPA}}) + \text{segnali_esterni} + \sigma_v \cdot W_v(t)$; stima empirica Kuppens et al. (2010): $\gamma_v \approx 0.40\text{-}0.65/\text{ora}$ (alta regolazione) vs $\gamma_v \approx 0.10\text{-}0.25/\text{ora}$ (bassa regolazione, depressione). Neuroanatomia amigdala: BLA $\rightarrow$ ICN (gate inibitorio GABAergico) $\rightarrow$ CeA; proiezioni CeA $\rightarrow$ PAG (dPAG per fuga/lotta; vlPAG per freezing) $\rightarrow$ PVN (asse HPA, CRH $\rightarrow$ ACTH $\rightarrow$ cortisolo) $\rightarrow$ LC (noradrenalina, consolidazione mnestica emotiva) $\rightarrow$ DMV/NA (tono vagale cardiaco). Via rapida (12 ms): talamo $\rightarrow$ BLA; via lenta (25-50 ms): talamo $\rightarrow$ corteccia $\rightarrow$ BLA. SSRI come modulatori della precisione serotoninergica: $\gamma_v^{\{\text{post-SSRI}\}} > \gamma_v^{\{\text{baseline}\}}$. Benzodiazepine: $\theta_{\text{CeA}}^{\{\text{BZD}\}} = \theta_{\text{CeA}}^{\{\text{baseline}\}} + \Delta\theta_{\{\text{GABA}\}} > 0$; meccanismo di tolleranza: down-regulation recettori GABA_A.

Monografia XXII: I giochi evolutivi della neurobiologia. Ricombinazione VDJ: diversità combinatoria $D_{\text{BCR}} \approx 10^{18}\text{-}10^{20}$. Diversità HLA: > 30.000 alleli; selezione dipendente dalla frequenza negativa $dp_i/dt = p_i \cdot (w_i(p) - \bar{w}(p))$ con w_i funzione decrescente di p_i (Red Queen). Studio Meyerhans et al. (1989): evoluzione virale HIV per sfuggire a HLA-B57:01. Modello Lotka-Volterra CTL-virus: $dT/dt = \lambda - d \cdot T - \beta \cdot V \cdot T$; $dI/dt = \beta \cdot V \cdot T - \delta \cdot I - k \cdot E \cdot I$; $dV/dt = p \cdot I - c \cdot V$; $dE/dt = q \cdot E \cdot I - d_E \cdot E$; frequenze di oscillazione $\omega \approx \sqrt{(k \cdot q \cdot I \cdot E^* - (\delta + k \cdot E^*)^2)/4}$. Modello quasispecies HIV-CTL: $dv_j/dt = r_j \cdot v_j \cdot (1 - \epsilon_j \cdot E) + \sum_k \mu\{k_j\} \cdot v_k - d \cdot v_j$. Ipotesi del cervello sociale (Dunbar, 1992): $\log(N_{\text{sociale}}) = a + b \cdot \log(N_Q)$, $b \approx 3.8$, $N_{\text{Dunbar}} \approx 148$. Struttura gerarchica delle reti sociali: $n_k \approx 3^k \cdot n_0$. Valore di Shapley: $\varphi_i(v) = \sum_{\{S \subseteq N \setminus \{i\}\}} \frac{|S|! \cdot (n - |S| - 1)!}{n!} \cdot [v(S \cup \{i\}) - v(S)]$.

Monografia XXIII: La verifica biologica. Frequenza errori DNA: $f_{\text{finale}} = f_{\text{pol}} \times r_{\text{proof}} \times r_{\text{MMR}} \approx 10^{-5} \times 10^{-2} \times 10^{-3} = 10^{-10}$. UPR: IRE1 α $\rightarrow$ splicing non convenzionale XBP1s (rimozione introne di 26 nt) + RIDD; PERK $\rightarrow$ fosforilazione eIF2 α (S51) $\rightarrow$ inibizione traduzione globale + ATF4 $\rightarrow$ CHOP; ATF6 $\rightarrow$ taglio S1P/S2P $\rightarrow$ ATF6f (fattore trascrizionale). Decisione adattamento vs. apoptosi: dipendente dalla cinetica del segnale, non dalla sola ampiezza. Autofagia: complesso ULK1 attivato da AMPK (stress energetico) e inibito da mTORC1 (nutrienti); complesso PI3K-Beclin1 $\rightarrow$ produzione PI3P $\rightarrow$ nucleazione fagoforo; sistemi di coniugazione ATG12-ATG5-ATG16L1 e LC3-PE; recettori cargo p62, OPTN, NDP52 (motivo LIR per legame a LC3-II). Mitofagia PINK1-Parkin: accumulo PINK1 su mitocondri con $\Delta\Psi$ ridotto $\rightarrow$ attivazione Parkin $\rightarrow$ ubiquitinazione OMM $\rightarrow$ autofagia selettiva. Telomeri: $L_n = L_0 - \Delta l \cdot n$ con $\Delta l \approx 100$ bp/divisione; $n_{\text{max}} \approx 100$ divisioni; limite di Hayflick ≈ 50 divisioni osservato. Telomerasi: TERT + TERC (template 3'-AAUCCC-5'). SASP: secrezione di IL-6, IL-8, TNF- α , MMP da cellule senescenti.

Monografia XXIV: Il corpo come indeterminazione. Chiasmo di Merleau-Ponty (*Le Visible et l'Invisible*, 1968): la Chair come medium comune del toccante e del toccato; reversibilità strutturale ma incompleta — la prima persona non coincide mai interamente con la terza. Neuroni specchio (Rizzolatti et al., 1992): corrispondenza funzionale tra azione eseguita e azione osservata nell'area F5 del macaco; sistema specchio umano (HMS): attivazione sovrapposta durante esecuzione e osservazione di azioni in BA 44/45 e lobulo parietale inferiore BA 40. Embodied simulation (Gallese, 2001): $\text{rappresentazione}_{S(a_O)} \approx \text{rappresentazione}_{S(a_S)}$ — comprensione attraverso risonanza motoria, non inferenza teorica. Imitazione neonatale (Meltzoff-Moore, 1977): intercorporeità filogeneticamente preformata. Monismo anomalo di Davidson: ogni evento mentale è identico a un evento fisico (monismo) ma le proprietà mentali non si riducono alle proprietà fisiche (anomalismo). Divario esplicativo (Levine): la distanza tra la descrizione fisica completa e la descrizione fenomenica è strutturale, non contingente.

Monografia I

I limiti della rappresentazione neurale

Il teorema di Minsky-Papert, la geometria dello spazio delle attivazioni e il problema del binding come impossibilità strutturale

I. La domanda che precede tutte le altre

Prima ancora di chiedersi come un sistema biologico pensi, ragioni o decida, occorre affrontare una questione più elementare e, in un certo senso, più radicale: in che cosa consiste rappresentare qualcosa? La neurobiologia moderna ha accumulato decenni di osservazioni sull'architettura corticale, sulla selettività delle singole cellule nervose, sulla struttura dei circuiti locali e sulle connessioni a lungo raggio tra aree. Eppure la domanda sulla natura della rappresentazione neurale rimane aperta in un senso che non è meramente tecnico, bensì strutturale: esistono proprietà del mondo che nessuna architettura di neuroni può rappresentare, indipendentemente da quante cellule siano disponibili e da quanto fitte siano le loro connessioni?

La risposta, che questa monografia sviluppa con la piena rigore formale che il tema esige, è affermativa — e il percorso verso quella risposta rivela qualcosa di sorprendente: i limiti della rappresentazione neurale non sono accidentali difetti di implementazione, bensì manifestazioni biologiche di impossibilità matematiche che appartengono alla stessa famiglia del teorema di incompletezza di Gödel e del teorema di Rice. Così come nessun sistema formale può dimostrare la propria consistenza, nessun sistema di unità di elaborazione a soglia può separare tutte le classi di pattern senza una struttura geometrica che eccede le capacità di qualsiasi strato singolo di elaborazione; e così come nessun algoritmo può verificare proprietà semantiche non banali di programmi arbitrari, nessun meccanismo neurale noto può integrare informazione distribuita in rappresentazioni unitarie senza pagare un prezzo in forma di vincoli temporali e di capacità che producono errori sistematici osservabili.

Il percorso di questa monografia attraversa tre territori distinti ma connessi. Il primo è il territorio della geometria della separabilità lineare, dove il teorema di Minsky-Papert fornisce il risultato fondamentale sull'espressività dei sistemi a singolo strato. Il secondo è il territorio della geometria delle rappresentazioni distribuite, dove la struttura degli spazi di attivazione neurali viene analizzata con gli strumenti della geometria differenziale e della teoria dell'informazione. Il terzo, e più profondo, è il territorio del problema del binding — la questione di come un sistema di unità di elaborazione parallele e distribuite riesca a costruire rappresentazioni unitarie di oggetti complessi, una questione

che si rivela non risolvibile in linea generale senza meccanismi espliciti di coordinazione che introducono a loro volta nuovi limiti strutturali.

II. Il teorema di Minsky-Papert: geometria della separabilità

II.1 Il percettrone come classificatore lineare

Il percettrone, introdotto da Frank Rosenblatt nel 1958 e formalizzato da Minsky e Papert nella monografia omonima del 1969, è il modello matematico più semplice di unità di elaborazione neurale. Un percettrone con n ingressi è una funzione

$$f: \mathbb{R}^n \rightarrow \{-1, +1\}$$

definita da

$$f(x) = \text{sign}(\sum_i w_i x_i - \theta)$$

dove $w = (w_1, \dots, w_n) \in \mathbb{R}^n$ è il vettore dei pesi sinaptici, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ è il vettore degli ingressi, e $\theta \in \mathbb{R}$ è la soglia di attivazione. Convenzionalmente si introduce un ingresso aggiuntivo $x_0 = 1$ con peso $w_0 = -\theta$, riducendo la funzione alla forma omogenea $f(x) = \text{sign}(w \cdot x)$.

La superficie di separazione del percettrone è l'iperpiano $\{x : w \cdot x = 0\}$ in $\mathbb{R}^n$. Il percettrone classifica il punto x come appartenente alla classe positiva se $w \cdot x > 0$, alla classe negativa altrimenti. Questa struttura geometrica immediata rivela il limite fondamentale: il percettrone può separare soltanto insiemi linearmente separabili.

Definizione (separabilità lineare). Due insiemi $S_+ \subset \mathbb{R}^n$ e $S_- \subset \mathbb{R}^n$ sono linearmente separabili se esiste un iperpiano $H = \{x : w \cdot x = 0\}$ tale che per ogni $x \in S_+$ si ha $w \cdot x > 0$ e per ogni $x \in S_-$ si ha $w \cdot x < 0$.

Il teorema di convergenza del percettrone, dimostrato da Rosenblatt (1962) e reso più preciso da Novikoff (1963), garantisce che la regola di apprendimento del percettrone converge in un numero finito di passi se e soltanto se le classi sono linearmente separabili.

Teorema (convergenza del percettrone, Novikoff 1963). Sia un dataset di training $\{(x_i, y_i)\}_{i=1}^M$ con $x_i \in \mathbb{R}^n$ e $y_i \in \{-1, +1\}$. Supponiamo che le classi siano linearmente separabili con margine $\gamma >$

0, ossia che esista w^* con $\|w^*\| = 1$ tale che $y_i(w^* \cdot x_i) \geq \gamma$ per ogni i . Sia $R = \max_i \|x_i\|$. Allora la regola di aggiornamento

$$w_{+1} = w + y_i x_i \text{ (quando } y_i(w \cdot x_i) \leq 0 \text{)}$$

commette al più $(R/\gamma)^2$ errori prima di convergere a un classificatore privo di errori.

Il bound $(R/\gamma)^2$ dipende dal rapporto tra il raggio dei dati e il margine di separazione: dati ben separati (grande γ) con ampiezza contenuta (piccolo R) conducono a convergenza rapida. Quando la separabilità lineare non è garantita, l'algoritmo non termina.

II.2 Il problema dell'XOR e la sua generalizzazione

L'esempio paradigmatico dell'inespressività del perceptrone singolo è il problema logico dello XOR (exclusive or), che Minsky e Papert usano come caso fondamentale. Lo XOR ha quattro punti in $\mathbb{R}^2$:

$$(0,0) \rightarrow -1, (0,1) \rightarrow +1, (1,0) \rightarrow +1, (1,1) \rightarrow -1$$

Questi quattro punti non sono linearmente separabili: qualsiasi retta nel piano divide i quattro punti in modo tale che almeno una delle due classi non è confinata in un unico semipiano. La dimostrazione elementare procede per assurdo: se esistesse un iperpiano $w_1x_1 + w_2x_2 = \theta$ che separa le classi, allora sommando le disuguaglianze per i punti positivi e sottraendo quelle per i punti negativi si ottiene una contraddizione aritmetica immediata. Nessuna retta — e più in generale nessun iperpiano in $\mathbb{R}^2$ — può separare $\{(0,1), (1,0)\}$ da $\{(0,0), (1,1)\}$.

La generalizzazione di Minsky e Papert va ben oltre questo esempio. Essi introducono la nozione di predicato su insiemi di punti in $\mathbb{R}^n$ e studiano la classe di predicati che un perceptrone con un numero limitato di connessioni per unità di input può calcolare. Il risultato fondamentale concerne i predicati di parità:

Teorema (Minsky-Papert, 1969). Il predicato di parità su n variabili binarie — che assegna il valore positivo quando il numero di variabili uguali a 1 è pari — richiede un perceptrone con almeno n connessioni per unità di input. Non esiste alcun perceptrone con un numero di connessioni sub-lineare in n che possa calcolare il predicato di parità.

Questo risultato colpisce perché il predicato di parità è computazionalmente elementare (calcolabile in tempo $O(n)$) ma strutturalmente inaccessibile ai perceptron con connettività locale. La dimostrazione sfrutta argomenti di simmetria: i predicati simmetrici rispetto a permutazioni delle variabili di input richiedono che il

perceptrone abbia accesso globale a tutte le variabili, il che, con connettività locale, è impossibile.

II.3 La dimensione di Vapnik-Chervonenkis e la capacità geometrica

La caratterizzazione completa dell'espressività di una classe di classificatori richiede la nozione di dimensione di Vapnik-Chervonenkis (VC), introdotta nel celebre articolo del 1971.

Definizione (frantumazione). Un insieme di m punti $\{x_1, \dots, x\} \subset \mathbb{R}^n$ è frantumato dalla classe di funzioni $\mathcal{H}$ se per ogni etichettatura $y : \{x_1, \dots, x\} \rightarrow \{-1, +1\}$ esiste una funzione $h \in \mathcal{H}$ tale che $h(x_i) = y_i$ per ogni i .

Definizione (dimensione VC). La dimensione VC di $\mathcal{H}$, indicata $d_{VC}(\mathcal{H})$, è la cardinalità del massimo insieme di punti che $\mathcal{H}$ può frantumare.

Per la classe dei perceptron lineari in $\mathbb{R}^n$ (classificatori della forma $\text{sign}(w \cdot x + b)$ con $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$), la dimensione VC vale esattamente $n + 1$. La dimostrazione si articola in due parti:

Lato superiore: per ogni insieme di $n + 2$ punti in $\mathbb{R}^n$, esiste un'etichettatura non frantumabile da perceptron lineari. Questo segue dal lemma di Radon, che afferma che ogni insieme di $n + 2$ punti in $\mathbb{R}^n$ può essere partizionato in due sottoinsiemi i cui involucri convessi si intersecano; la coppia di involucri che si intersecano non può essere separata da un iperpiano.

Lato inferiore: esiste un insieme di $n + 1$ punti che il perceptrone può frantumare. Si sceglie l'insieme formato da $\{e_1, e_2, \dots, e, 0\}$ dove e_i è l' i -esimo vettore della base canonica. Per ogni etichettatura di questi $n + 1$ punti, si costruisce esplicitamente un iperpiano separatore.

La dimensione VC è la caratterizzazione esatta della capacità di memorizzazione m , attraverso il teorema fondamentale della teoria PAC, la caratterizzazione della complessità di campionamento richiesta per la generalizzazione.

Teorema (complessità di campionamento PAC). Una classe $\mathcal{H}$ con $d_{VC}(\mathcal{H}) = d$ è PAC-apprendibile con complessità di campionamento

$$m(\epsilon, \delta) = \Theta(d/\epsilon + \log(1/\delta)/\epsilon)$$

Nel caso del perceptrone, $d = n + 1$: la complessità di campionamento cresce linearmente con la dimensione dell'input.

II.4 Il lemma di Cover e la probabilità di separabilità

Un risultato complementare, dovuto a Cover (1965), fornisce la prospettiva probabilistica sulla separabilità lineare. Data una configurazione di m punti in posizione generale in $\mathbb{R}^n$ (ossia nessun insieme di $n + 1$ punti è affinementemente dipendente), la probabilità che un'assegnazione di etichette casuale sia linearmente separabile è

$$C(m, n) = (1/2^{m-1}) \cdot \sum_{k=0}^n C(m-1, k)$$

dove $C(m-1, k)$ è il coefficiente binomiale. Questa formula, nota come funzione di capacità di Cover, ha il comportamento asintotico

$$C(m, n) \rightarrow 1 \text{ per } m < 2n \quad C(m, n) \rightarrow 0 \text{ per } m > 2n$$

La transizione di fase avviene intorno a $m = 2n$: quando il numero di punti da classificare supera il doppio della dimensione dell'input, la probabilità che una classificazione casuale sia linearmente separabile crolla verso zero esponenzialmente. Questo risultato ha un'interpretazione biologica precisa: un'area corticale con n neuroni di proiezione non può classificare in modo affidabile collezioni di pattern con più di $2n$ elementi, indipendentemente dall'architettura dei pesi.

III. Rappresentazioni distribuite e la geometria dello spazio delle attivazioni

III.1 Dalla separabilità lineare alla geometria delle varietà

Il teorema di Minsky-Papert stabilisce i limiti di una singola unità lineare a soglia. I sistemi biologici reali implementano, tuttavia, trasformazioni composte attraverso molteplici strati di elaborazione non lineare. La domanda pertinente diventa: quali strutture geometriche emergono nello spazio delle attivazioni di reti multi-strato, e quali proprietà fondamentali non possono essere catturate da nessuna architettura finita?

L'ipotesi della varietà (manifold hypothesis), formulata esplicitamente da Bengio, Courville e Vincent (2013) e già implicita nel lavoro di Tenenbaum, de Silva e Langford (2000), propone che i dati di interesse cognitivo — immagini di oggetti naturali, sequenze di parlato, pattern di movimento — non siano distribuiti uniformemente nello spazio ad alta dimensione degli ingressi sensoriali, bensì concentrati in vicinanza di varietà di dimensione intrinseca molto più bassa. Un'immagine naturale

di 256×256 pixel occupa uno spazio di dimensione 65.536, ma lo spazio delle immagini naturali significative ha dimensione intrinseca stimata tra 100 e 200 per scene complesse.

La struttura formale è la seguente. Sia $\mathcal{M} \subset \mathbb{R}^D$ una varietà differenziabile di dimensione $d \ll D$, e sia X una variabile aleatoria con distribuzione concentrata in un intorno tubolare di $\mathcal{M}$ di raggio ϵ . Una rete neurale profonda $F : \mathbb{R}^D \rightarrow \mathbb{R}^d$ che apprende la struttura di $\mathcal{M}$ produce una rappresentazione

$$z = F(x) \in \mathbb{R}^d$$

con la proprietà che F sia localmente omeomorfa a $\mathcal{M}$: nell'intorno di ogni punto $x \in \mathcal{M}$, F si comporta come una carta locale della varietà. Questa struttura è precisamente quella di un autoencoder che minimizza la distanza di ricostruzione

$$\mathcal{L} = E_x[\|x - G(F(x))\|^2]$$

dove $G : \mathbb{R}^d \rightarrow \mathbb{R}^D$ è il decodificatore. Il risultato fondamentale è che, nel limite di capacità sufficiente e dati abbondanti, la composizione $G \circ F$ converge all'operatore di proiezione sulla varietà $\mathcal{M}$ (Fefferman, Mitter e Narayanan, 2016).

III.2 Il principio di codifica sparsa e la sovrapposizione

Il cervello biologico non implementa rappresentazioni localiste — dove ogni concetto corrisponde a un singolo neurone dedicato — bensì rappresentazioni distribuite, in cui ogni concetto attiva un pattern di attività su un ampio numero di cellule, e ogni cellula partecipa a molti concetti distinti. Questa struttura è formalizzata dal principio di codifica sparsa, introdotto da Olshausen e Field (1996) nel contesto della corteccia visiva primaria.

Siano $x \in \mathbb{R}^n$ il vettore di input (ad esempio un frammento di immagine naturale) e $A \in \mathbb{R}^{n \times m}$ un dizionario di m atomi (con $m \gg n$, sistema sovracompleto). La codifica sparsa cerca la rappresentazione

$$z^* = \operatorname{argmin}_{\{z \in \mathbb{R}^m\}} \|x - Az\|^2 + \lambda \|z\|_1$$

dove il termine $\|z\|_1 = \sum_i |z_i|$ penalizza il numero di coefficienti non nulli e promuove la sparsità. Olshausen e Field dimostrano che quando il dizionario A è appreso dai dati di immagini naturali ottimizzando la sparsità, le colonne di A assomigliano ai campi recettivi dei neuroni della corteccia visiva primaria V1 dell'animale: bande orientate, localizzate in frequenza, che rispondono a bordi e texture locali.

Il meccanismo alla base della corrispondenza tra codifica sparsa e fisiologia è il seguente. Un segnale x generato da una sorgente sparsa — con poche caratteristiche attive simultaneamente — è più comprimibile nella base A che ne cattura le componenti indipendenti. La corteccia visiva, evolvendosi in un ambiente dove le immagini naturali hanno struttura statistica sparsa (un campo visivo naturale ha poche strutture salienti per ogni momento), ha sviluppato un'architettura che implementa approssimativamente la codifica sparsa.

La formalizzazione della distribuzione delle immagini naturali è cruciale. Ruderman e Bialek (1994) mostrano che le correlazioni spaziali nelle immagini naturali seguono la legge di potenza

$$C(\Delta x) = \langle I(x)I(x + \Delta x) \rangle \propto |\Delta x|^{-\alpha}$$

con $\alpha \approx 2$, indicando invarianza di scala. La trasformata di Fourier di C è lo spettro di potenza

$$S(f) \propto 1/f^2$$

La struttura $1/f^2$ dello spettro implica che le immagini naturali sono statisticamente auto-simili: le loro proprietà statistiche sono invarianti rispetto a cambiamenti di scala. Questa struttura è ciò che la codifica sparsa apprende, producendo basi che sono locali, orientate e selettive in frequenza.

III.3 La teoria della superposizione come limite strutturale

Il principio di codifica sparsa implica che ogni neurone partecipi a molteplici rappresentazioni. Quando il numero di concetti da rappresentare supera il numero di neuroni disponibili, la rete deve implementare la superposizione: la co-attivazione di pattern sovrapposti per concetti distinti. Questa struttura, analizzata formalmente da Elhage et al. (2022) nel contesto dei transformer, produce interferenze che costituiscono un limite strutturale della rappresentazione distribuita.

Definizione (superposizione). Un sistema di n neuroni implementa la superposizione di $m > n$ feature quando la matrice di embedding $W \in \mathbb{R}^{n \times m}$ (che mappa ogni feature nel suo pattern di attivazione) soddisfa $W^T W \approx I_m$ (le feature sono approssimativamente ortogonali nello spazio delle attivazioni) anche con $n < m$.

Questo è possibile soltanto approssimativamente, con un errore che cresce con il grado di sovra-completezza m/n . L'errore di interferenza tra la feature i e la feature j è

$$\|(\mathbf{W}^T \mathbf{W})_{ii}\|^2 \approx (n/m)(1 - n/m)$$

nel caso di embedding casuali uniformi. La probabilità di interferenza catastrofica — che la presenza della feature j attivi erroneamente la feature i — è controllata dalla sparsità delle feature: se la frequenza di attivazione di ciascuna feature è $\varphi \ll 1$, l'interferenza media è

$$\langle \text{errore} \rangle \approx \varphi \cdot m/n$$

Il limite fondamentale è che per $\varphi \rightarrow 0$ (feature molto rare) il sistema può rappresentare un numero di feature m che scala come $n/\varphi \gg n$, mentre per $\varphi \rightarrow 1$ (feature sempre attive) il sistema non può rappresentare più di n feature senza interferenze catastrofiche.

La connessione biologica è precisa: la corteccia cerebrale opera in un regime di sparsità elevata (stime sperimentali danno $\varphi \approx 0.01\text{--}0.05$ in varie aree corticali), il che le consente di rappresentare un numero di concetti superiore al numero di neuroni di un fattore dell'ordine di 20–100. Ma questo regime di sparsità è esso stesso un vincolo: richiede meccanismi inibitori attivi (i neuroni inibitori di tipo GABAergico che costituiscono circa il 20% di tutte le cellule corticali) per mantenere il livello di attività basso, e questi meccanismi hanno costi metabolici e temporali propri.

IV. Il problema del binding: l'impossibilità della coerenza distribuita

IV.1 La formulazione di Treisman e la teoria dell'integrazione delle feature

Il problema del binding — come il sistema visivo integra informazioni distribuite su feature distinte (colore, forma, orientamento, moto) in una rappresentazione unitaria di un oggetto — fu posto nella forma moderna da Anne Treisman e Garry Gelade (1980) attraverso la teoria dell'integrazione delle feature (Feature Integration Theory, FIT).

La FIT propone una distinzione fondamentale tra:

- feature elementari (colore, orientamento, luminanza, movimento) che vengono elaborate pre-attentivamente, in parallelo e senza effort, in mappe separate di feature;
- oggetti complessi (congiunzioni di feature) la cui percezione richiede un processo attentivo seriale che "colla" le feature in un insieme coerente.

L'evidenza sperimentale più convincente è quella della ricerca visiva (visual search). Quando si cerca un target definito da una singola feature elementare tra distrattori — ad esempio una barra rossa tra barre verdi — i tempi di ricerca sono indipendenti dal numero di distrattori (ricerca in parallelo). Quando il target è definito da una congiunzione di feature — ad esempio una barra rossa orientata verticalmente tra barre rosse orizzontali e barre verdi verticali — i tempi di ricerca crescono linearmente con il numero di distrattori (ricerca seriale).

La formalizzazione quantitativa è la seguente. Siano $F_1, F_2, \dots, F_k$ le k mappe di feature elementari. Ogni mappa $F_i : \mathbb{R}^2 \rightarrow V_i$ assegna a ogni posizione del campo visivo un valore nella varietà di valori della feature i (ad esempio $V_{\text{colore}} = S^1$ per il tono cromatico, $V_{\text{orientamento}} = \mathbb{RP}^1$ per l'orientamento). La rappresentazione di un oggetto complesso richiede la congiunzione dei valori di tutte le feature in una medesima posizione:

$$\text{obj}(p) = (F_1(p), F_2(p), \dots, F_k(p))$$

Il problema è che questa congiunzione non è disponibile senza un meccanismo esplicito di accesso alle singole mappe di feature nella posizione p — un meccanismo che Treisman identifica con l'attenzione spaziale. In assenza di attenzione, le feature "galleggiano" libere nelle proprie mappe e possono essere erroneamente ricombinate, generando oggetti illusori (illusory conjunctions): il sistema percettivo può vedere un cerchio rosso e un triangolo blu quando sono presenti un cerchio blu e un triangolo rosso, perché le feature si sono legate al posto sbagliato.

IV.2 La soluzione temporale e i suoi limiti: l'ipotesi della sincronizzazione

La risposta neurofisiologica più influente al problema del binding è l'ipotesi della sincronizzazione oscillatoria, sviluppata da Wolf Singer e Charles Gray a partire dalla fine degli anni Ottanta. L'idea centrale è che i neuroni che codificano feature appartenenti allo stesso oggetto si sincronizzino nella loro attività — oscillino nella stessa fase nella banda gamma (30–80 Hz) — mentre i neuroni che codificano feature di oggetti distinti rimangono asincroni.

La formalizzazione matematica dell'ipotesi della sincronizzazione si basa sulla teoria delle oscillazioni accoppiate. Sia $\varphi_i(t)$ la fase dell'oscillazione del neurone i al tempo t . L'equazione del moto per un sistema di oscillatori accoppiati di Kuramoto è

$$d\varphi_i/dt = \omega_i + (K/N) \sum \sin(\varphi - \varphi_i)$$

dove ω_i è la frequenza naturale dell'oscillatore i , K è la costante di accoppiamento, e N è il numero di oscillatori. La transizione di fase del modello di Kuramoto avviene al valore critico

$$K_c = 2 / (\pi g(0))$$

dove $g(\omega)$ è la distribuzione delle frequenze naturali. Per $K > K_c$ la popolazione esibisce sincronizzazione parziale; per $K \rightarrow \infty$ tutti gli oscillatori convergono alla stessa fase.

Il parametro di ordine di Kuramoto

$$r(t) = |(1/N) \sum \exp(i\varphi(t))|$$

misura il grado di sincronizzazione: $r = 0$ per oscillatori completamente asincroni, $r = 1$ per oscillatori perfettamente sincronizzati.

Tuttavia, il meccanismo della sincronizzazione come soluzione al problema del binding incontra due obiezioni formali di rilievo che ne limitano la generalità.

Prima obiezione: la capacità di separazione. Se la sincronizzazione distingue oggetti diversi attraverso differenze di fase, quanti oggetti può separare simultaneamente? Per oscillatori che devono mantenere distinguibilità di fase con differenza angolare minima $\Delta\varphi$, il numero massimo di oggetti separabili è

$$N_{\max} = \lfloor 2\pi / \Delta\varphi \rfloor$$

Poiché la risoluzione temporale della fase nelle oscillazioni gamma è dell'ordine di 1–2 ms e il ciclo gamma ha durata 15–33 ms, la differenza di fase minima rilevabile è $\Delta\varphi \approx \pi/10$, dando $N_{\max} \approx 5-6$ oggetti. Questo corrisponde notevolmente con la capacità della memoria di lavoro visiva (circa 4 oggetti, secondo le stime sperimentali di Cowan 2001 e Luck e Vogel 1997), ma implica un limite assoluto che non dipende dalla potenza computazionale del sistema.

Seconda obiezione: la scalabilità del binding temporale. Il binding attraverso sincronizzazione richiede che i neuroni in aree diverse (V1, V4, IT, area frontale) sincronizzino le proprie oscillazioni. La velocità di propagazione dei segnali nervosi lungo le connessioni corticali a lungo raggio è dell'ordine di 1–10 m/s, il che implica ritardi di trasmissione tra aree distanti di 5–50 ms. Per la banda gamma a 40 Hz, un ritardo di 25 ms corrisponde esattamente a mezza lunghezza d'onda — ciò che dovrebbe essere sincronizzato diventa invece anti-sincronizzato. I sistemi nervosi reali compensano parzialmente questo problema attraverso la selezione della fase nelle connessioni a lungo raggio

(Bastos et al., 2015), ma questo compensatore introduce ulteriore complessità e ulteriori limiti.

IV.3 La formulazione algebrica del binding e il teorema di impossibilità

La struttura del problema del binding può essere formalizzata in termini algebrici in modo da rendere esplicita la natura della impossibilità che esso comporta. Questa formulazione, elaborata nei suoi contorni più precisi da Smolensky (1990) e poi da Plate (1995), rivela che il binding è strutturalmente un problema di rappresentazione di strutture relazionali in spazi vettoriali.

Siano $a, b, c, d \in \mathbb{R}^n$ le rappresentazioni di quattro feature elementari (ad esempio: rosso, blu, cerchio, quadrato). La rappresentazione di due oggetti — un cerchio rosso e un quadrato blu — richiede di codificare due legami:

$$\text{colore}(\text{cerchio}) = \text{rosso} \text{ e } \text{colore}(\text{quadrato}) = \text{blu}$$

In un sistema di rappresentazione distributa lineare, ogni oggetto occupa lo stesso spazio vettoriale $\mathbb{R}^n$ in cui sono rappresentate le feature. Non esiste un meccanismo lineare che consenta di recuperare le associazioni in modo affidabile quando più oggetti sono rappresentati simultaneamente.

Il prodotto tensoriale fornisce la soluzione matematica esatta: la coppia (cerchio, rosso) è rappresentata dal tensore $a \otimes b \in \mathbb{R}^{n \times n}$, con n^2 dimensioni. Ma il tensore non è computazionalmente conveniente: con $n = 1000$ neuroni per feature, ogni oggetto richiede $n^2 = 10^6$ unità, e k oggetti richiedono $k \cdot n^2$ unità — una crescita quadratica intrattabile.

Le rappresentazioni vettoriali distribuite (Holographic Reduced Representations di Plate, Random Indexing, e le rappresentazioni iperdimensionali di Kanerva) offrono un compromesso: approssimano il prodotto tensoriale attraverso un'operazione di legame (binding) $\circledast : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ che mappa coppie in vettori della stessa dimensione, preservando l'ortogonalità approssimativa. La binding operation più semplice è la convoluzione circolare (Plate 1995):

$$(a \circledast b) = \sum a \cdot b_{\{k-j \bmod n\}}$$

Il recupero è approssimativo: dato il vettore legato $a \circledast b$ e il vettore a , si può recuperare $\tilde{b} \approx b$ attraverso la correlazione

$$\tilde{b} = \tilde{a} \circledast^\dagger (a \circledast b)$$

dove $\bar{a}$ è la versione invertita di a e $\circledast$ è la correlazione circolare. Il recupero non è esatto ma approssimato, con un errore che dipende dal numero di legami simultaneamente presenti nello spazio.

Teorema (impossibilità del binding esatto in spazi fissi). Non esiste un'operazione $\circledast : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ che soddisfi simultaneamente le tre proprietà seguenti:

- (i) Compositività: il vettore $(a \circledast b) \circledast c$ rappresenta la struttura (a, b, c) .
- (ii) Separabilità: da $(a \circledast b)$ e a si può recuperare b esattamente.
- (iii) Capacità: k oggetti distinti possono essere legati con errore di recupero zero per $k > 1$.

Dimostrazione. Per la proprietà (ii), l'operazione $\circledast$ deve essere invertibile a sinistra: esiste $\circledast^{-1}$ tale che $b = a \circledast^{-1} (a \circledast b)$. Per la proprietà (iii), il vettore $v = (a_1 \circledast b_1) + (a_2 \circledast b_2) + \dots + (a_k \circledast b_k)$ deve permettere il recupero di b_i da a_i . Applicando $\circledast^{-1}$ al primo termine: $a_1 \circledast^{-1} v = b_1 + \sum_{i \geq 2} a_1 \circledast^{-1} (a_i \circledast b_i)$. Per il recupero esatto, si richiede $a_1 \circledast^{-1} (a_i \circledast b_i) = 0$ per ogni $i \geq 2$, il che implica che $a_i \circledast b_i$ sia nel nucleo dell'operatore $a_1 \circledast^{-1}$. Ma lo spazio dei vettori in $\mathbb{R}^n$ ha dimensione finita n , e il nucleo di un operatore lineare ha co-dimensione pari al rango dell'operatore. Per $k > n + 1$ oggetti, non è possibile che tutti i vettori $a_i \circledast b_i$ ($i \geq 2$) stiano nel nucleo di ogni operatore di recupero: si giunge alla contraddizione che lo spazio $\mathbb{R}^n$ dovrebbe avere una dimensione infinita. $\square$

Il teorema mostra che il binding esatto di più oggetti simultanei in uno spazio di dimensione fissa è impossibile: qualsiasi sistema di rappresentazione distribuita deve accettare errori che crescono con il numero di oggetti rappresentati simultaneamente. I sistemi biologici gestiscono questa impossibilità attraverso meccanismi complementari: l'attenzione seriale che riduce $k \approx 1$ nell'unità di tempo, la memoria di lavoro che mantiene $k \leq 4$ oggetti in una rappresentazione altamente interferente, e l'ipocampo che fornisce una struttura sparsa, a pattern separati, per la codifica episodica di eventi complessi.

V. La selettività neurale come proiezione su varietà

V.1 I campi recettivi come kernel di elaborazione

L'organizzazione gerarchica della corteccia visiva — dal corpo genicolato laterale (LGN) attraverso V1, V2, V4, fino alla corteccia inferotemporale (IT) — implementa una sequenza di trasformazioni

che progressivamente costruiscono rappresentazioni invarianti di oggetti. La formalizzazione di questa gerarchia in termini di kernel e spazi di feature offre la struttura matematica per analizzarne i limiti.

Un neurone con campo recettivo è caratterizzato da una funzione di risposta

$$r = f(x * k + b)$$

dove $x \in \mathbb{R}^{H \times W}$ è l'immagine di input (H righe e W colonne), $k \in \mathbb{R}^{h \times w}$ è il kernel (il profilo del campo recettivo), $*$ indica la convoluzione locale, b è la soglia, e f è la funzione di attivazione non lineare (tipicamente la funzione di rettificazione ReLU o la funzione sigmoide).

Per i neuroni semplici di V1 (Hubel e Wiesel, 1959, 1968), il kernel k ha la forma di un filtro di Gabor orientato:

$$k(x, y; \sigma_x, \sigma_y, \theta, f_0, \varphi) = \exp(-(x'^2 / 2\sigma_x^2 + y'^2 / 2\sigma_y^2)) \cdot \cos(2\pi f_0 x' + \varphi)$$

dove $x' = x \cos \theta + y \sin \theta$ e $y' = -x \sin \theta + y \cos \theta$ sono le coordinate ruotate dell'angolo θ rispetto all'asse orizzontale, σ_x e σ_y definiscono l'estensione del campo recettivo lungo e perpendicolarmente all'orientamento preferito, f_0 è la frequenza spaziale preferita, e φ è la fase. Lo spazio dei parametri (θ, f_0, φ) definisce la "mappa" dei campi recettivi in V1: ogni punto dell'area corticale è caratterizzato da un particolare valore di θ , f_0 e φ , e l'insieme di tutte le cellule di V1 campiona densamente lo spazio tridimensionale di questi parametri.

La misura di Gabor ottimizza il guadagno di informazione per segnali naturali nella rappresentazione sparsa: i filtri di Gabor sono le funzioni che massimizzano simultaneamente la localizzazione spaziale e quella in frequenza, nel senso della relazione di Heisenberg per segnali:

$$\sigma_x \cdot \sigma_f \geq 1/(4\pi)$$

dove σ_x è la larghezza del campo recettivo e σ_f è la larghezza della banda passante in frequenza. Il prodotto minimo è raggiunto esattamente dalle funzioni di Gabor, che sono dunque le rappresentazioni ottimali per segnali spaziali nel senso del principio di incertezza tempo-frequenza.

V.2 Invarianza e il ruolo delle trasformazioni di gruppo

Man mano che si sale nella gerarchia visiva (da V1 verso IT), le risposte neurali diventano progressivamente invarianti rispetto a trasformazioni dello stimolo: traslazione, scala, rotazione, illuminazione. La struttura

matematica dell'invarianza è quella della teoria dei gruppi applicata alle trasformazioni dello stimolo.

Sia G il gruppo delle trasformazioni di interesse (ad esempio il gruppo $SE(2)$ delle isometrie piane per la traslazione e la rotazione, o il gruppo di scala $\mathbb{R}_+$ per le trasformazioni di dimensione). Un neurone con risposta $r(x)$ è invariante rispetto a G se $r(g \cdot x) = r(x)$ per ogni $g \in G$ e ogni stimolo x , dove $g \cdot x$ denota l'azione di g su x .

L'equivarianza è una proprietà più debole che preserva la struttura trasformativa: $r(g \cdot x) = g \cdot r(x)$. La convoluzione spaziale è equivariante rispetto alle traslazioni: traslare l'immagine di input traslazione identicamente la mappa delle attivazioni, senza distorcerne la forma.

La costruzione sistematica di rappresentazioni invarianti per un gruppo G passa attraverso il pooling sulle orbite. L'orbita di x sotto G è

$$O_G(x) = \{g \cdot x : g \in G\}$$

La media dell'operatore di risposta sulle orbite

$$\bar{r}(x) = \int_G r(g \cdot x) dg$$

è invariante per costruzione. Ma il pooling sulle orbite distrugge l'informazione sulla trasformazione specifica che ha prodotto lo stimolo osservato — il principio dell'invarianza-discriminabilità (DiCarlo e Cox, 2007): più un sistema è invariante rispetto a G , meno può discriminare tra stimoli che si trovano sulla stessa orbita.

La gerarchia visiva corticale realizza un compromesso graduato: i livelli inferiori ($V1$, $V2$) sono altamente selettivi ma poco invarianti; i livelli superiori (IT) sono altamente invarianti ma meno selettivi per i dettagli fini. Questo compromesso è necessario — non un difetto di progettazione — perché l'invarianza totale e la discriminabilità totale sono mutuamente esclusive per trasformazioni non banali.

V.3 La geometria di Riesgold-Shepard e la legge universale di generalizzazione

Roger Shepard (1987) ha proposto un risultato che egli stesso chiama "legge universale di generalizzazione": la probabilità che un organismo generalizzi una risposta appresa per uno stimolo x a uno stimolo y è una funzione decrescente della distanza tra x e y nello spazio psicologico, e specificamente è una funzione esponenziale di quella distanza:

$$P(\text{generalizzazione: } x \rightarrow y) = \exp(-c \cdot d(x, y))$$

dove c è una costante che dipende dall'importanza della generalizzazione e $d(x, y)$ è la distanza tra x e y in uno spazio metrico appropriato.

Shepard dimostra che questa forma esponenziale è universale nel senso che emerge necessariamente da qualsiasi sistema di rappresentazione ottimale della struttura probabilistica delle categorie naturali, indipendentemente dai dettagli del sistema. La dimostrazione sfrutta il principio bayesiano di massima entropia: se si assume che le categorie naturali siano "naturali" nel senso di occupare regioni convesse di estensione casuale nello spazio degli stimoli, allora la probabilità di generalizzazione ottimale è esattamente la funzione esponenziale della distanza.

La legge di Shepard ha un corollario formale importante per i limiti della rappresentazione: poiché la distanza psicologica è una metrica, lo spazio psicologico è necessariamente uno spazio metrico. Ma le strutture concettuali umane violano sistematicamente la metricità: la somiglianza percepita non è simmetrica (un volto che assomiglia alla Monna Lisa non è percepito nella stessa misura in cui la Monna Lisa assomiglia a quel volto), e la disuguaglianza triangolare è violata per triplette di stimoli che attraversano confini categoriali (Tversky, 1977). Queste violazioni mostrano che la rappresentazione mentale non è semplicemente uno spazio metrico di dimensione fissa, ma ha una struttura dipendente dal contesto e dalla salienza categoriale che nessuno spazio metrico fisso può catturare.

VI. I limiti computazionali del sistema nervoso biologico

VI.1 La velocità di elaborazione e i vincoli temporali

Il sistema nervoso biologico opera sotto vincoli temporali che costituiscono limiti strutturali indipendenti dalla teoria della separabilità e del binding. Un neurone corticale tipico produce un potenziale d'azione ogni 5–100 ms (frequenza di scarica 10–200 Hz), con un periodo refrattario assoluto di circa 1–2 ms durante il quale non può produrre un secondo potenziale d'azione. La propagazione del potenziale d'azione lungo un assone mielinizzato (diametro 10–20 μm) avviene alla velocità di 50–100 m/s; lungo un assone non mielinizzato (0.5–2 μm) alla velocità di 0.5–2 m/s.

Poiché il cervello umano ha dimensioni lineari dell'ordine di 15 cm, la latenza di trasmissione tra regioni distanti è

$$\tau \approx 0.15 \text{ m} / (1 \text{ m/s}) \approx 150 \text{ ms per fibre non mielinizzate}$$

e

$$\tau \approx 0.15 \text{ m} / (50 \text{ m/s}) \approx 3 \text{ ms per fibre mielinizzate rapide}$$

La percezione visiva consapevole richiede circa 150–200 ms dall'onset dello stimolo. In questo intervallo, le informazioni devono attraversare i 10 strati della gerarchia visiva (retina $\rightarrow$ LGN $\rightarrow$ V1 $\rightarrow$ V2 $\rightarrow$ V4 $\rightarrow$ IT $\rightarrow$ PFC), con latenza di 10–20 ms per transizione sinaptica. Questo lascia, in media, meno di 20 ms per ogni stadio di elaborazione: un periodo durante il quale ogni neurone può produrre al massimo 2–4 potenziali d'azione. In conseguenza, le decisioni percettive devono basarsi su codici neurali molto brevi, spesso con un singolo spike per neurone, il che pone in questione la possibilità di utilizzare il rate coding (la frequenza di scarica come misura dell'intensità dello stimolo) per l'elaborazione rapida (Thorpe e Imbert, 1989).

VI.2 Il paradosso della codifica temporale e la degenerazione

La risposta al problema della scarsità di tempo è la codifica temporale: invece di trasmettere l'intensità attraverso la frequenza di scarica, i neuroni trasmettono l'informazione attraverso la latenza del primo spike dopo l'onset dello stimolo (rank-order coding di Thorpe et al., 1996). In questo schema, la risposta neurale al tempo relativo del primo spike t_i rispetto all'onset dello stimolo è inversamente correlata con la preferenza del neurone per quello stimolo: i neuroni che preferiscono di più lo stimolo scaricano prima.

La capacità di trasmissione del codice a primo spike è formalmente stimabile con la teoria dell'informazione. Se la distribuzione dei tempi di primo spike ha densità $p(t)$, la quantità di informazione trasmessa per spike è

$$H[T] = -\int p(t) \log p(t) dt$$

Per una distribuzione uniforme sull'intervallo $[t_{\min}, t_{\max}]$, questa vale $\log(t_{\max} - t_{\min})$. Con $t_{\max} - t_{\min} \approx 100 \text{ ms}$ e risoluzione temporale $\approx 1 \text{ ms}$, la capacità per spike è circa 7 bit — notevolmente superiore ai circa 2–3 bit trasmissibili attraverso un codice di frequenza in 20 ms con jitter neurale tipico.

Il paradosso della degenerazione si pone quando si considera che neuroni diversi con campi recettivi sovrapposti rispondono allo stesso stimolo con latenze simili: come si distingue l'identità del neurone che ha scaricato per primo (l'informazione sul contenuto della feature che quel neurone rappresenta) dal tempo assoluto del primo spike (l'informazione sull'intensità della feature)? Questa degenerazione è intrinseca al codice a rango temporale e richiede che il sistema sia in grado di stimare sia l'identità della fonte che il tempo relativo — due misure non separabili in linea di principio senza informazione a priori sulla risposta media di ciascun neurone.

VII. La coscienza come problema di rappresentazione: il caso della prosopagnosia

Il limite della rappresentazione neurale trova la sua manifestazione clinica più rivelatrice nei disturbi selettivi del riconoscimento visivo. La prosopagnosia — l'incapacità di riconoscere i volti familiari nonostante la conservazione della vista e dell'intelligenza generale — illustra la granularità della rappresentazione corticale: il riconoscimento dei volti è implementato in regioni specifiche (il giro fusiforme del volto, FFA, nell'area BA37) che possono essere danneggiate selettivamente. Il danno lascia intatte le capacità di riconoscere oggetti non biologici, corpi umani senza volto e scene visive complesse.

Questo fenomeno pone il problema del binding in una luce diversa: la rappresentazione dei volti non è soltanto il binding di feature (occhi, naso, bocca), bensì una rappresentazione olistica in cui le relazioni metriche tra le feature sono codificate come un pattern globale, non compositamente. L'interferenza olistica di Tanaka e Farah (1993) mostra che la parte superiore di un volto viene riconosciuta più accuratamente quando è presentata nel contesto della parte inferiore del volto originale che quando è presentata isolata: il contesto trasforma la rappresentazione.

La struttura formale di questa olicità è quella di una rappresentazione basata sul modello completo (template-based) piuttosto che su feature elementari: la risposta dell'FFA è meglio predetta dalla correlazione della trasformata wavelet dell'immagine con un template appreso che dalla risposta a feature locali. Questo schema di rappresentazione è radicalmente diverso dalla decomposizione in feature locali e indipendenti di V1, e pone il problema di come i due sistemi di rappresentazione — quello locale-compositivo e quello

olisco-relazionale — vengano integrati. Il binding tra questi due livelli di rappresentazione è un problema aperto che le soluzioni temporali (sincronizzazione) e spaziali (pooling) non affrontano nella sua generalità.

VIII. Conclusioni: la rappresentazione come confine strutturale

Il percorso di questa monografia ha attraversato i risultati fondamentali sulla natura e i limiti della rappresentazione neurale. Il teorema di Minsky-Papert stabilisce che la separabilità lineare è il confine geometrico invalicabile per i sistemi a singolo strato, con implicazioni che si estendono alle architetture multi-strato attraverso la struttura delle varietà. La teoria della sovrapposizione mostra che la rappresentazione distribuita di concetti multipli in spazi di dimensione fissa introduce interferenze strutturali che non possono essere eliminate senza ridurre la capacità. Il problema del binding rivela un'impossibilità fondamentale: il legame esatto di feature multiple in rappresentazioni coerenti richiede risorse che crescono almeno quadraticamente con il numero di feature, e qualsiasi soluzione approssimativa nel medesimo spazio dimensionale introduce errori irriducibili.

Questi limiti non sono difetti contingenti che una tecnologia superiore potrebbe superare: sono manifestazioni biologiche di vincoli matematici dello stesso ordine dei teoremi di incompletezza e di Rice. La rappresentazione neurale è strutturalmente limitata non perché il cervello sia un cattivo ingegnere, ma perché i problemi che deve risolvere — la classificazione invariante di oggetti complessi in tempo reale con risorse metaboliche finite — toccano i confini di ciò che qualsiasi sistema di elaborazione fisica può fare.

Comprendere questi confini è il prerequisito per comprendere sia l'architettura del cervello reale — come si è evoluta per abitare quel territorio con soluzioni ingenue ma necessariamente approssimate — sia le possibilità e i limiti dei sistemi artificiali che aspirano a replicarne le prestazioni. L'evoluzione ha avuto quattro miliardi di anni per esplorare lo spazio delle soluzioni; la sua risposta ai confini della rappresentazione è il tema che le monografie successive svilupperanno.

Monografia II

I limiti dell'ottimizzazione evolutiva

Il teorema fondamentale di Fisher, la deriva genetica, i paesaggi adattativi e la struttura delle impossibilità evolutive

I. L'evoluzione come algoritmo di ricerca e le sue impossibilità

L'evoluzione per selezione naturale è, nella formulazione più astratta, un algoritmo di ricerca nello spazio dei genotipi: esplora lo spazio delle varianti ereditabili, valuta ciascuna variante attraverso la sua idoneità riproduttiva, e amplifica le varianti con idoneità superiore alla media della popolazione. Questa struttura ha indotto una lunga tradizione di pensiero a considerare l'evoluzione come un ottimizzatore — lento, contingente, privo di teleologia esplicita, ma efficace nel produrre soluzioni adattate ai problemi posti dall'ambiente.

Questa visione, per quanto utile come euristica, è formalmente scorretta in sensi precisi che questa monografia si propone di enunciare con rigore. L'evoluzione non è un ottimizzatore nel senso matematico del termine per almeno tre ragioni strutturali distinte che corrispondono a tre classi di impossibilità.

La prima è la contingenza storica del processo: la selezione naturale ottimizza localmente rispetto al paesaggio adattativo corrente, e il paesaggio è esso stesso in mutamento continuo per effetto delle co-evoluzioni, dei cambiamenti ambientali e delle innovazioni evolutive degli altri organismi. Non esiste un "ottimo globale" stabile verso cui il processo converga.

La seconda è la stocasticità intrinseca della deriva genetica: anche in popolazioni con pressione selettiva forte e costante, il campionamento casuale delle generazioni introduce fluttuazioni nelle frequenze alleliche che possono fissare varianti deleterie ed eliminare varianti favorevoli, con probabilità che dipendono dalla dimensione della popolazione e che sono strutturalmente irriducibili.

La terza — e più profonda — è il teorema No Free Lunch applicato all'evoluzione: non esiste un algoritmo evolutivo che superi un algoritmo di ricerca casuale quando mediato su tutti i possibili paesaggi adattativi. La selezione naturale è efficace soltanto perché i paesaggi

adattativi reali hanno struttura — non sono casuali — ma questa struttura non è garantita in linea di principio e in certi regimi si dissolve.

Il percorso di questa monografia sviluppa questi tre temi con la strumentazione matematica appropriata: il teorema fondamentale della selezione naturale di Fisher e le sue qualificazioni formali, la teoria della deriva genetica di Kimura, la struttura dei paesaggi adattativi NK di Kauffman, e il teorema No Free Lunch di Wolpert e Macready. Il filo rosso è la dimostrazione che l'evoluzione — come l'apprendimento automatico, come la logica formale — abita un territorio delimitato da impossibilità strutturali, e che la comprensione di queste impossibilità è il prerequisito per comprendere i sistemi evolutivi reali.

II. Il teorema fondamentale della selezione naturale: enunciato preciso e condizioni di validità

II.1 La formalizzazione di Fisher

Ronald Fisher enunciò nel 1930 un risultato che definì il "teorema fondamentale della selezione naturale" e che equiparò, in audacia esplicativa, al secondo principio della termodinamica. La versione originale afferma:

Il tasso di aumento dell'idoneità di qualsiasi organismo in qualsiasi momento è uguale alla sua varianza genetica dell'idoneità in quel momento.

La formalizzazione matematica precisa richiede alcune definizioni. Sia $p = (p_1, p_2, \dots, p_n)$ il vettore delle frequenze alleliche di n alleli a un singolo locus, con $p_i \geq 0$ e $\sum_i p_i = 1$. L'idoneità assoluta dell'individuo con genotipo (i, j) in un organismo diploide è W_{ij} , e la frequenza del genotipo (i, j) in una popolazione in equilibrio di Hardy-Weinberg è $p_i p_j$ (per $i \neq j$ si prende $2p_i p_j$). L'idoneità marginale dell'allele i è

$$w_i = \sum_j p_j W_{ij}$$

e l'idoneità media della popolazione è

$$\bar{w} = \sum_i p_i w_i = \sum_i \sum_j p_i p_j W_{ij}$$

L'equazione di Fisher per il cambiamento della frequenza allelica in tempo discreto (generazioni) è

$$\Delta p_i = p_i(w_i - \bar{w}) / \bar{w}$$

In tempo continuo, assumendo che le generazioni si sovrappongano:

$$dp_i/dt = p_i(w_i - \bar{w})$$

Questa equazione, nota come equazione di selezione, descrive il cambiamento delle frequenze alleliche per effetto della selezione differenziale.

Il teorema fondamentale afferma che il tasso di cambiamento dell'idoneità media per effetto della selezione — e soltanto per effetto della selezione, non includendo i cambiamenti ambientali o le interazioni genotipo-ambiente — è uguale alla varianza additiva dell'idoneità. La varianza additiva è la componente della varianza totale dell'idoneità che dipende linearmente dalle frequenze alleliche e che può essere aumentata dalla selezione.

La varianza additiva in un sistema a un locus con idoneità W_i è

$$V_A = \sum_i p_i (w_i - \bar{w})^2$$

Il teorema di Fisher afferma, nella versione precisa corretta da Price (1972) e Ewens (1989):

$$d\bar{w}/dt|_{e \square e z_{i0} \square e} = V_A / \bar{w}$$

Cruciale è la qualificazione: il termine $d\bar{w}/dt|_{e \square e z_{i0} \square e}$ non è la variazione totale di $\bar{w}$ nel tempo, ma soltanto la componente dovuta alle variazioni di frequenza allelica tenendo fissi i valori di W_i . La variazione totale include contributi degli effetti ambientali mutevoli, delle mutazioni, della deriva genetica e delle co-evoluzioni con altri organismi — e questi contributi possono essere negativi e di ampiezza arbitraria.

II.2 La dimostrazione attraverso il prezzo dell'equazione di Price

George Price (1970, 1972) fornì la derivazione più generale e formalmente pulita del teorema di Fisher attraverso quella che oggi porta il suo nome.

L'equazione di Price descrive il cambiamento di qualsiasi carattere z in qualsiasi processo evolutivo. Sia $\bar{W}_i = W_i/\bar{W}$ l'idoneità relativa dell'individuo i (numero di discendenti relativamente alla media). Sia z_i il valore del carattere nell'individuo i , e $\bar{z}$ la media del carattere nei discendenti di i (il carattere può cambiare dalla generazione parentale a quella filiale per trasmissione imperfetta). L'equazione di Price è:

$$\bar{W} \Delta \bar{z} = \text{Cov}(\bar{W}, \bar{z}) + E[\bar{W} \Delta \bar{z}]$$

dove $\text{Cov}(W, z) = E[Wz] - E[W]E[z] = E[Wz] - W\bar{z}$ è la covarianza tra idoneità e carattere nella generazione parentale, e $E[W\Delta z] = E[W(z' - z)]$ è il termine di trasmissione che cattura i cambiamenti del carattere all'interno di ogni lignaggio dalla generazione parentale a quella filiale.

Il primo termine, $\text{Cov}(W, z)/W$, rappresenta l'effetto della selezione: se la covarianza tra idoneità e carattere è positiva, gli individui con idoneità superiore hanno valori superiori del carattere, e la selezione aumenta $\bar{z}$. Il secondo termine, $E[W\Delta z]/W$, rappresenta l'effetto di tutto ciò che non è selezione: mutazioni, deriva, ricombinazione, sviluppo.

Per recuperare il teorema fondamentale di Fisher, si pone $z = W$ (il carattere è l'idoneità stessa). Allora

$$W\Delta\bar{w} = \text{Cov}(W, W) + E[W\Delta W]$$

Il primo termine $\text{Cov}(W, W) = \text{Var}(W)$ è la varianza dell'idoneità. Il secondo termine $E[W\Delta W]$ cattura i cambiamenti dell'idoneità all'interno dei lignaggi — che possono essere negativi se l'ambiente cambia o se le co-evoluzioni riducono l'idoneità degli individui con idoneità inizialmente alta.

La versione rigorosa del teorema di Fisher è dunque:

$$\Delta\bar{w} \big|_{e \square e Z_{io} \square e} = \text{Var}(W) / W$$

con la qualificazione che questo è soltanto il contributo della selezione alla variazione totale di $\bar{w}$. Il risultato è immediatamente analogico al secondo principio della termodinamica nella formulazione di Clausius: l'entropia (varianza dell'idoneità) cresce sempre localmente per effetto della selezione, ma l'entropia totale (compreso l'effetto dell'ambiente) può diminuire.

II.3 I limiti immediati del teorema

Il teorema fondamentale di Fisher è spesso citato come prova che l'evoluzione ottimizza l'idoneità: se la varianza genetica dell'idoneità è sempre positiva, l'idoneità media cresce sempre per effetto della selezione. Ma questa interpretazione è formalmente scorretta in almeno tre sensi.

Primo limite: l'esaurimento della varianza. Il teorema afferma che la selezione fa crescere $\bar{w}$ a un tasso proporzionale a V_A . Ma la selezione stessa riduce V_A aumentando le frequenze degli alleli favorevoli e riducendo quelle degli alleli sfavorevoli. In una popolazione in cui la selezione converge all'allele ottimale (un unico allele fissa), $V_A \rightarrow 0$ e il processo si ferma. Il teorema garantisce crescita finché c'è varianza, ma

non garantisce che la varianza si mantenga né che l'ottimo raggiunto sia un massimo globale piuttosto che un punto di sella locale.

Secondo limite: la selezione su caratteri poligenetici. Per caratteri determinati da molti loci indipendenti — come la statura, l'intelligenza, la resistenza alle malattie — la varianza additiva totale è la somma delle varianze contribute da ciascun locus:

$$V_A = \sum_l V_A^{(l)}$$

Il teorema si applica a ciascun locus separatamente, ma la dinamica dell'evoluzione multigenica dipende in modo cruciale dalle correlazioni tra i loci (disequilibrio di linkage), dalla dominanza e dall'epistasi — interazioni non lineari tra alleli a loci diversi. In presenza di epistasi, la covarianza tra idoneità e carattere non è linearmente collegata alla varianza additiva, e il teorema perde la sua forma semplice.

Terzo limite: la stazionarietà del paesaggio. Il teorema assume implicitamente che i valori di idoneità W_i siano costanti nel tempo. In ambienti reali, $W_i = W_i(t)$ varia continuamente per effetto dei cambiamenti climatici, dell'evoluzione dei predatori, dei parassiti e delle specie competitive. Il secondo termine dell'equazione di Price — $E[W\Delta W]$ — può essere arbitrariamente negativo, e in certi casi la popolazione si allontana dall'ottimo corrente più rapidamente di quanto la selezione possa avvicinarla al nuovo ottimo.

III. La deriva genetica e il ruolo del caso

III.1 La teoria neutralista di Kimura

Motoo Kimura (1968, 1983) propose una teoria dell'evoluzione che per un decennio sembrò in radicale contrasto con la visione darwiniana: la maggior parte dei cambiamenti evolutivi alle molecole (sostituzioni di nucleotidi nel DNA, cambiamenti di amminoacidi nelle proteine) è adattativamente neutra — non ha effetti sulla fitness — e si fissa nelle popolazioni non per selezione positiva ma per deriva genetica casuale.

La deriva genetica è la fluttuazione delle frequenze alleliche dovuta al campionamento casuale delle generazioni. In una popolazione di dimensione N , la frequenza allelica p in una generazione è il risultato del campionamento di $2N$ geni (per un organismo diploide) dalla generazione precedente. Se la frequenza in una generazione è p , la frequenza nella generazione successiva ha distribuzione

$$P(k \text{ alleli su } 2N) = C(2N, k) p^k (1-p)^{(2N-k)}$$

con media $E[p'] = p$ e varianza

$$\text{Var}(p') = p(1-p) / (2N)$$

La varianza è inversamente proporzionale alla dimensione della popolazione: popolazioni piccole fluttuano molto, popolazioni grandi poco. In una popolazione finita, un allele neutro (con idoneità identica all'allele alternativo) si fissa con probabilità uguale alla sua frequenza iniziale. Per un allele di nuova mutazione presente in una singola copia, la probabilità di fissazione è

$$P_{\text{fissazione}} = 1/(2N)$$

Il tempo medio di fissazione di un allele neutro, condizionato alla fissazione, è

$$T_{\text{fissazione}} \approx 4N \text{ generazioni}$$

Questo risultato fondamentale — il tasso di sostituzione molecolare per alleli neutri è uguale al tasso di mutazione μ , indipendentemente dalla dimensione della popolazione — è il risultato chiave della teoria neutrale: il tempo di fissazione scala con N , ma la frequenza con cui la mutazione avviene scala con N allo stesso modo (N mutazioni per locus per generazione in una popolazione di dimensione N), e le due dipendenze si cancellano esattamente.

Il tasso di sostituzione molecolare netto è dunque

$$k = 2N \cdot \mu / (2N) = \mu$$

La molecola evolve a una velocità (in sostituzioni per generazione) uguale al tasso di mutazione, indipendentemente dalla dimensione della popolazione. Questa previsione, confermata con sorprendente precisione dall'analisi delle sequenze proteiche e del DNA, costituisce la cosiddetta "orologeria molecolare" e fornisce il fondamento della filogenesi molecolare.

III.2 La probabilità di fissazione con selezione: la formula di Kimura

Il caso non neutrale — un allele con vantaggio selettivo s rispetto all'allele alternativo — è descritto dalla formula di fissazione di Kimura (1962), derivata dall'equazione di diffusione per il processo di Wright-Fisher.

Sia p la frequenza iniziale dell'allele favorevole. La probabilità di fissazione in una popolazione di dimensione N con vantaggio selettivo s e dominanza h è

$$P(p, N, s) = [1 - \exp(-4Nsp)] / [1 - \exp(-4Ns)]$$

Per una singola copia mutante ($p = 1/(2N)$) e per $Ns \gg 1$:

$P_{\text{fissazione}} \approx 2s$ (per alleli dominanti, $h = 1$) $P_{\text{fissazione}} \approx s$ (per alleli recessivi, $h = 0$) $P_{\text{fissazione}} \approx 2hs$ (per alleli con dominanza h)

Il risultato fondamentale è che anche un allele fortemente favorevole ($s = 0.01$) ha soltanto il 2% di probabilità di fissazione quando appare in una singola copia. Il 98% delle mutazioni vantaggiose viene perso per deriva prima di diventare abbastanza frequenti da essere efficacemente "viste" dalla selezione.

Questo risultato — la probabilità di fissazione è dell'ordine di $2s$ per alleli ad alto vantaggio — ha un corollario importante per la comprensione dei limiti evolutivi: l'evoluzione non può sfruttare varianti vantaggiose di piccolo effetto in popolazioni piccole. Per $N \cdot s \ll 1$, la probabilità di fissazione dell'allele vantaggioso non differisce significativamente da quella dell'allele neutro ($1/(2N)$), e la selezione è "invisibile" alla deriva. La grandezza $Ne \cdot s$ — il prodotto della dimensione efficace della popolazione per il coefficiente di selezione — è la quantità adimensionale che determina il regime evolutivo: per $Ne \cdot s \gg 1$ domina la selezione, per $Ne \cdot s \ll 1$ domina la deriva.

III.3 Il ratchet di Müller e l'accumulo irreversibile di mutazioni

Hermann Joseph Müller (1964) osservò che nelle popolazioni asessuate, la classe degli individui con il minor numero di mutazioni deleterie non può essere ripristinata una volta persa. Questo meccanismo — noto come "ratchet di Müller" — porta all'accumulo irreversibile di mutazioni deleterie nelle popolazioni asessuate di piccola dimensione.

La formalizzazione è la seguente. Sia n il numero di individui con esattamente k mutazioni deleterie, e sia n_0 il numero di individui nella classe con zero mutazioni (la "classe zero"). In assenza di ricombinazione, la classe zero può essere ricostruita soltanto da se stessa attraverso la riproduzione. La frequenza della classe zero è

$$f_0 = n_0 / N$$

dove N è la dimensione totale della popolazione. La probabilità che la classe zero si estingua per deriva in una generazione è dell'ordine di

$\exp(-n_0)$ per n_0 piccolo. Ogni volta che la classe zero si estingue, viene sostituita dalla classe $k = 1$ (individui con una sola mutazione), che diventa la nuova "classe zero". Il processo è irreversibile: il ratchet scatta in avanti ma non indietro.

La velocità del ratchet — il numero di scatti per generazione — dipende dalla dimensione della popolazione N , dal tasso di mutazione μ e dalla diminuzione di idoneità per mutazione s . Haigh (1978) derivò l'approssimazione per la velocità del ratchet:

$$v \approx \mu \exp(-U/s) / T$$

dove $U = N\mu$ è il numero di mutazioni per genoma per generazione, s è l'effetto di ciascuna mutazione sulla fitness, e T è il tempo medio di estinzione della classe zero. Per $U \gg s$ (alte mutazioni, bassa selezione per mutazione), il ratchet è molto veloce; per $U \ll s$ (basse mutazioni, alta selezione) il ratchet quasi non scatta.

Il ratchet di Müller è considerato uno dei principali vantaggi evolutivi del sesso: la ricombinazione sessuale mescola i genomi parentali, permettendo la ricostruzione di classi a bassa mutazione attraverso la combinazione di gameti complementari. In termini formali, la ricombinazione permette che la classe $k = 0$ venga ricostituita da due individui della classe $k = 1$, purché le loro mutazioni siano in posizioni diverse. Il sesso è dunque, in questa prospettiva, una soluzione evolutiva a un problema di impossibilità computazionale: l'accumulo irreversibile di errori in sistemi di copia sequenziale senza meccanismi di correzione inter-copia.

IV. I paesaggi adattativi e la topologia dell'ottimizzazione evolutiva

IV.1 Il concetto di paesaggio adattativo e le sue difficoltà formali

Sewall Wright (1932) introdusse la metafora del paesaggio adattativo: uno spazio in cui ogni genotipo corrisponde a un punto, la dimensionalità è uguale al numero di loci, e l'altezza in ogni punto rappresenta l'idoneità del genotipo corrispondente. La selezione naturale corrisponde a un processo di "salita sulla collina" (hill-climbing) che tende a portare la popolazione verso i massimi locali dell'idoneità.

La metafora è intuitivamente potente ma formalmente problematica per diverse ragioni. Primo, lo spazio dei genotipi ha dimensionalità astronomica: un genoma con L loci diploidi e k alleli per locus ha $k^{(2L)}$ possibili genotipi, un numero che per valori biologicamente realistici ($L = 10^6$ loci, $k = 2$ alleli) vale $2^{(2 \cdot 10^6)}$ — un numero infinitamente superiore agli atomi dell'universo osservabile. Lo spazio dei genotipi non può essere "esplorato" ma soltanto campionato attraverso la storia evolutiva di una linea genetica.

Secondo, l'idoneità di un genotipo non è una proprietà intrinseca del genotipo stesso ma dipende dall'ambiente e, crucialmente, dalla composizione della popolazione: in un sistema coevolutivo (predatore-preda, parassita-ospite, maschio-femmina), l'idoneità di un genotipo dipende dalle frequenze di tutti gli altri genotipi nella popolazione. Il paesaggio adattativo è quindi un oggetto dipendente dal tempo e dalla composizione della popolazione — non una superficie statica.

Terzo, la topologia dello spazio dei genotipi — le relazioni di vicinanza tra genotipi — dipende dall'architettura del genoma. Due genotipi sono "vicini" se differiscono per poche mutazioni puntiformi; ma la struttura delle reti metaboliche, degli enhancer trascrizionali e dei moduli di sviluppo fa sì che mutazioni in loci distanti possono avere effetti interattivi (epistatici) molto maggiori di mutazioni in loci vicini. La distanza metrica nello spazio dei genotipi non è la distanza rilevante per l'evoluzione.

IV.2 Il modello NK di Kauffman

Stuart Kauffman (1993) propose il modello NK come formalizzazione della struttura dei paesaggi adattativi. Nel modello NK, un genoma è una stringa binaria $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_N)$ di N loci con valori $\{0, 1\}$. L'idoneità del genotipo è

$$W(\sigma) = (1/N) \sum_i w_i(\sigma_i, \sigma_{\{i_1\}}, \sigma_{\{i_2\}}, \dots, \sigma_{\{i_K\}})$$

dove w_i è una funzione della variabile i -esima e delle sue K interazioni con altri K loci (scelti casualmente), e i valori di w_i sono estratti indipendentemente da una distribuzione uniforme su $[0, 1]$. Il parametro K misura il grado di epistasi: per $K = 0$ (nessuna interazione), ogni locus contribuisce indipendentemente e il paesaggio ha un unico massimo globale (il genotipo con tutti i valori ottimali per ciascun locus); per $K = N-1$ (interazioni complete), ogni locus interagisce con tutti gli altri e il paesaggio è completamente casuale.

Le proprietà del paesaggio NK dipendono crucialmente da K .

Per $K = 0$, il paesaggio ha un unico massimo globale e il processo di hill-climbing converge garantito a quel massimo. L'altezza del massimo globale è $E[W_{\max}] = 1 - 1/(N+1)$.

Per $K = N-1$, il paesaggio è equivalente a un paesaggio casuale (ogni genotipo ha idoneità indipendente da ogni altro). Il numero di massimi locali scala come

$$E[M] \approx 2^N / (N+1)$$

ed è esponenzialmente grande. Ogni massimo locale attrae un bacino di attrazione di dimensioni medie $2^N / E[M] \approx N+1$ — un intorno molto piccolo rispetto all'intero spazio dei genotipi. Il hill-climbing si arena in uno dei massimi locali con alta probabilità e l'ottimo globale è irraggiungibile in tempo polinomiale.

Per K intermedi, il numero di massimi locali scala come

$$E[M] \approx 2^N / (K+2)$$

e cresce esponenzialmente con K . La profondità media dei minimi locali — la distanza in termini di mutazioni necessarie per uscire dal bacino di attrazione — scala come

$$d_{\text{trap}} \approx N / (K+1)$$

Per K moderato ($K \approx 5-10$), il paesaggio ha una struttura di "bacini di attrazione" separati da barriere di altezza moderata, simile alla struttura dei vidrio di spin in fisica statistica. Il hill-climbing si arena in massimi locali subottimali, e la probabilità di trovare il massimo globale attraverso mutazioni singole decresce esponenzialmente con N per $K > 1$.

IV.3 La struttura delle barriere e il ruolo delle mutazioni neutrali

Una delle scoperte più importanti dell'evoluzione molecolare è che un gran numero di sostituzioni evolutive osservate sono neutre (Kimura) o quasi-neutre (King e Jukes, 1969). Questa neutralità ha un ruolo fondamentale nell'attraversamento delle barriere del paesaggio adattativo.

Maynard Smith (1970) propose il concetto di "rete di sequenze" per descrivere la struttura dello spazio dei genotipi: due sequenze sono connesse nella rete se differiscono per una sola sostituzione e hanno entrambe idoneità positiva (ossia sono "vitali"). La domanda è se la rete sia connessa — se cioè sia possibile passare da qualsiasi sequenza

funzionale a qualsiasi altra attraverso una sequenza di mutazioni singole che non riducano mai l'idoneità al di sotto di una soglia minima.

Fontana e Schuster (1998) studiarono questa questione per le molecole di RNA, dove l'idoneità è determinata dalla struttura secondaria (la struttura a clip delle molecole di RNA si può calcolare esattamente attraverso la programmazione dinamica). Il risultato sorprendente è che lo spazio delle sequenze di RNA è altamente connesso: quasi tutte le strutture secondarie possibili sono raggiungibili da quasi tutte le sequenze di partenza attraverso mutazioni singole neutre che attraversano la rete di sequenze neutre. Lo spazio delle sequenze è percorso da "reti neutrali" — insiemi di sequenze con la stessa struttura e quindi la stessa idoneità approssimativa — che si estendono attraverso l'intero spazio delle sequenze.

La struttura formale della rete neutrale può essere descritta come segue. Sia G lo spazio dei genotipi (stringhe di lunghezza N sull'alfabeto A) e $\varphi : G \rightarrow F$ la mappa genotipo-fenotipo (che mappa ogni genotipo al corrispondente fenotipo, con fenotipo = struttura per le molecole di RNA). La rete neutra del fenotipo f è l'insieme

$$N(f) = \{\sigma \in G : \varphi(\sigma) = f\}$$

Il diametro della rete neutra — la massima distanza (numero di mutazioni) tra due sequenze nella stessa rete — scala per molte strutture di RNA come

$$\text{diam}(N(f)) \approx \alpha N$$

con $\alpha \approx 0.6-0.8$ per le strutture tipiche: due sequenze della stessa rete neutrale distano mediamente il 60–80% della lunghezza della sequenza. Questo significa che la rete neutrale si estende per quasi tutto lo spazio delle sequenze, e percorrendola (attraverso mutazioni neutre) si possono raggiungere tutte le "vicinanze" dello spazio dove le altre strutture (altri fenotipi) sono accessibili attraverso mutazioni singole.

Il teorema biologicamente fondamentale è dunque:

Teorema (connettività delle reti neutrali, Fontana-Schuster 1998).

Per quasi tutte le strutture di RNA di lunghezza N , la rete neutrale $N(f)$ è connessa e ha diametro αN . La rete neutrale tocca un intorno ϵ -vicinanza di quasi ogni sequenza nello spazio delle sequenze, per $\epsilon = O(\log N)$.

L'implicazione evolutiva è che le mutazioni neutrali non sono irrilevanti: percorrendo la rete neutrale, la popolazione si muove attraverso lo spazio delle sequenze e accede a regioni da cui nuovi fenotipi adattativi (strutture diverse) sono raggiungibili attraverso mutazioni singole. La

neutralità è il meccanismo attraverso cui l'evoluzione attraversa le barriere del paesaggio adattativo — non saltandole, ma aggirando con l'equivalente topologico del percorso di minima resistenza.

V. Il teorema No Free Lunch e l'inefficacia dell'ottimizzazione senza struttura

V.1 Enunciato formale

Il teorema No Free Lunch (NFL) per gli algoritmi di ottimizzazione, dimostrato da Wolpert e Macready (1997), afferma che nessun algoritmo di ricerca/ottimizzazione è superiore a un altro quando i confronti vengono mediati su tutte le possibili funzioni obiettivo.

Sia F lo spazio di tutte le funzioni $f : X \rightarrow Y$ con X spazio finito di soluzioni candidate e Y spazio finito di valori. Sia A un algoritmo di ricerca che, dopo m valutazioni di f su punti $x_1, \dots, x_m \in X$ (scelti in base ai risultati delle valutazioni precedenti), produce una serie di valori $y_1 = f(x_1), \dots, y_m = f(x_m)$. Il comportamento dell'algoritmo A sulla funzione f è caratterizzato dalla distribuzione di probabilità dei valori osservati:

$$P(d_m^A \mid f, m, A)$$

dove $d_m^A = ((x_1, y_1), \dots, (x_m, y_m))$ è la storia delle m valutazioni.

Il teorema NFL afferma: per qualsiasi due algoritmi A_1 e A_2 , e per qualsiasi misura di performance Φ (ad esempio la qualità media del miglior valore trovato dopo m valutazioni):

$$\sum_{f \in F} \Phi(d_m^{A_1} \mid f) = \sum_{f \in F} \Phi(d_m^{A_2} \mid f)$$

La somma della performance sulla distribuzione uniforme di tutte le funzioni $f \in F$ è identica per qualsiasi algoritmo. In particolare, per la distribuzione uniforme su tutte le funzioni, la selezione naturale non è migliore della ricerca casuale.

V.2 La dimostrazione

La dimostrazione segue direttamente dalla simmetria dello spazio delle funzioni. Per qualsiasi algoritmo deterministico A che seleziona punti in base alla storia, e per qualsiasi permutazione π di Y , definire la funzione permutata $f^\pi(x) = \pi(f(x))$. L'algoritmo A che segue la storia $(x_1, y_1), \dots, (x_m, y_m)$ sulla funzione f segue la storia $(x_1, \pi(y_1)), \dots, (x_m, \pi(y_m))$ sulla funzione f^π — la sequenza di punti campionati è identica, perché

l'algoritmo non può distinguere tra f e f^π senza conoscenza a priori della struttura di π .

Poiché la distribuzione uniforme su F è chiusa rispetto alle permutazioni π di Y (permutare i valori di ogni funzione non cambia la distribuzione uniforme su F), la distribuzione dei valori osservati da A nella distribuzione uniforme su F è identica per ogni algoritmo A . Nessun algoritmo può essere sistematicamente migliore di un altro sulla distribuzione uniforme. $\square$

V.3 Implicazioni per l'evoluzione: la necessità della struttura ambientale

Il teorema NFL applicato all'evoluzione afferma che la selezione naturale come algoritmo di ottimizzazione non è superiore alla deriva genetica casuale quando mediata su tutti i possibili paesaggi adattativi. L'evoluzione funziona — produce adattamento — soltanto perché i paesaggi adattativi reali hanno struttura che non è uniforme e casuale.

La struttura rilevante dei paesaggi adattativi reali ha tre componenti principali.

Prima componente: la correlazione locale. In paesaggi biologici reali, genotipi vicini (che differiscono per poche mutazioni) tendono ad avere idoneità simili. Questa correlazione — assente nei paesaggi casuali del modello NK con $K = N-1$, presente nei paesaggi a bassa epistasi con K piccolo — permette al hill-climbing di fare progressi sistematici. La correlazione locale è quantificabile attraverso la lunghezza di correlazione del paesaggio:

$$\lambda = -1 / \ln \varrho(1)$$

dove $\varrho(d) = \text{Cov}[W(\sigma), W(\sigma')] / \text{Var}[W(\sigma)]$ è la correlazione dell'idoneità tra genotipi a distanza d mutazioni. Per paesaggi con $\lambda \gg 1$ (lunga correlazione), il hill-climbing è efficiente; per $\lambda \approx 1$ (breve correlazione), il hill-climbing non è migliore della ricerca casuale.

Seconda componente: la modularità. I genomi biologici sono modulari: i geni si raggruppano in pathway metabolici, circuiti regolatori e moduli di sviluppo che hanno effetti semi-indipendenti sull'idoneità. La modularità riduce la dimensione efficace dello spazio di ricerca da 2^N a $\sum_m 2^{(N_m)}$, dove N_m è la dimensione del modulo m , una riduzione esponenziale nel numero di moduli.

Terza componente: la gerarchicità del tempo evolutivo. Su scale di tempo evolutive brevi (migliaia di generazioni), la struttura del paesaggio è relativamente stabile e la selezione può fare progressi locali;

su scale di tempo lunghe (milioni di generazioni), il paesaggio si rimodella per effetto delle co-evoluzioni e dei cambiamenti ambientali, e le innovazioni evolutive — la comparsa di nuovi piani corporei, di nuovi metabolismi, di nuovi sistemi sensoriali — richiedono l'esplorazione di regioni dello spazio genotipico lontane dalla posizione corrente. Questa struttura gerarchica temporale del paesaggio consente all'evoluzione di combinare l'ottimizzazione locale a breve termine con l'esplorazione a lungo termine mediata dalla neutralità e dalla deriva.

VI. Gli spandrel e il Goodhart biologico: quando la selezione ottimizza il proxy sbagliato

VI.1 La critica di Gould e Lewontin

Stephen Jay Gould e Richard Lewontin (1979) pubblicarono un articolo destinato a diventare uno dei più influenti nella biologia del Novecento: "The Spandrels of San Marco and the Panglossian Paradigm". Il titolo si riferisce agli spandri della basilica di San Marco a Venezia — le superfici triangolari che si formano all'incrocio di archi — come metafora biologica. Gli spandri sono necessari dalla struttura architettonica, non sono stati "progettati" con uno scopo specifico, eppure vengono riempiti con mosaici bellissimi; da un punto di vista ingenuo si potrebbe essere tentati di spiegare gli spandri come "strutture progettate per contenere i mosaici".

La metafora biologica è quella dei caratteri evolutivi che non sono adattamenti diretti alla selezione ma conseguenze strutturali di vincoli di sviluppo, fisici o architettonici. L'argomento di Gould e Lewontin è che l'adattatismo — la tendenza a spiegare ogni carattere biologico come il prodotto della selezione naturale diretta su quel carattere — è metodologicamente sbagliato perché ignora i vincoli strutturali che determinano quali varianti sono possibili e quali no.

La formalizzazione rigorosa dell'argomento degli spandri richiede di distinguere tra due classi di caratteri:

Caratteri primari: caratteri su cui la selezione agisce direttamente perché hanno effetti diretti sull'idoneità. Esempio: la lunghezza del becco nei fringuelli di Darwin in relazione alla dimensione dei semi disponibili.

Caratteri secondari (spandri): caratteri che non vengono selezionati direttamente ma emergono come conseguenza necessaria della selezione

su caratteri primari o come conseguenze dei vincoli strutturali dello sviluppo. Esempio: il colore del sangue (rosso per l'emoglobina) non è stato selezionato perché il rosso è adattativo, ma è una conseguenza strutturale della composizione chimica dell'emoglobina che è stata selezionata per le sue proprietà di trasporto dell'ossigeno.

La struttura formale è quella di una mappa di sviluppo $\varphi : G \rightarrow P$ da genotipi a fenotipi — una mappa che non è una biiezione ma è multi-a-uno (molti genotipi producono lo stesso fenotipo) e non è onto (molti fenotipi potenziali non sono raggiungibili da nessun genotipo). Il dominio di φ nello spazio dei fenotipi — l'insieme dei fenotipi realizzabili $P^* = \text{Im}(\varphi)$ — è un sottoinsieme proprio dello spazio di tutti i fenotipi concepibili.

VI.2 Il Goodhart biologico: la selezione come cattivo ottimizzatore

Il parallelismo con la legge di Goodhart (discussa nella Monografia III del ciclo precedente) è strutturalmente preciso e illuminante. La legge di Goodhart afferma che quando una misura M diventa un obiettivo, cessa di essere una buona misura dell'obiettivo reale O — perché il sistema ottimizza M in modi che non ottimizzano O . Il meccanismo del Goodhart biologico è il seguente.

La selezione naturale opera su caratteri fenotipici osservabili (M): struttura morfologica, comportamento, fisiologia. Ma l'obiettivo evolutivo — se tale termine ha un senso — è l'idoneità inclusiva (O): il contributo di un individuo al pool genetico delle generazioni future, ponderato per la parentela. La selezione ottimizza il proxy fenotipico M , che nella maggior parte dei casi è correlato con O ma non identico ad esso.

Quando l'ambiente cambia, la correlazione tra M e O può dissolversi o invertirsi: caratteri che erano adattativi in un ambiente diventano neutri o maladattativi in un altro. Ma la selezione non "sa" questo: continua a ottimizzare il fenotipo corrente in relazione all'ambiente corrente, e può spingere la popolazione in direzioni che diminuiscono l'idoneità a lungo termine.

L'esempio classico è la selezione sessuale intrasessuale — la selezione dovuta alla competizione tra maschi per l'accesso alle femmine. La selezione favorisce i maschi più grandi, più aggressivi e con ornamenti più vistosi (coda del pavone, corna del cervo, colori brillanti), non perché questi caratteri siano ottimali per la sopravvivenza, ma perché sono indicatori di qualità genetica (ipotesi di Zahavi, 1975: handicap principle) o semplicemente perché piacciono alle femmine (Fisher runaway selection). Questi caratteri possono essere genuinamente

costosi in termini di sopravvivenza — la coda del pavone è un fardello per sfuggire ai predatori — ma vengono selezionati per i vantaggi riproduttivi che conferiscono.

La formalizzazione di Fisher (1930) della selezione runaway è elegante. Siano p la frequenza dell'allele che determina la preferenza delle femmine per le code lunghe, e q la frequenza dell'allele che determina la lunghezza della coda nei maschi. In presenza di correlazione genetica tra le femmine che preferiscono code lunghe e i maschi che hanno code lunghe (correlazione che si costruisce attraverso l'appaiamento assortativo), il sistema di equazioni accoppiato

$$\frac{dp}{dt} = \text{Var}_p \cdot (\text{cov}(p, w)/\bar{w}) + \text{cov}(p, \Delta q) \quad \frac{dq}{dt} = \text{Var}_q \cdot (\text{cov}(q, w)/\bar{w}) + \text{cov}(q, \Delta p)$$

ammette soluzioni instabili in cui p e q crescono entrambi in modo inarrestabile (runaway selection), anche quando la coda lunga riduce l'idoneità di sopravvivenza dei maschi. Il sistema "sfugge" al controllo dell'ottimizzazione adattativa ed entra in un regime di amplificazione positiva che può portare a caratteri esagerati lontani dall'ottimo adattativo.

VII. La coevoluzione come gioco e il Red Queen

VII.1 La dinamica coevolutiva preda-predatore

L'idoneità di un genotipo non è mai assoluta ma sempre relativa alla composizione genotipica della popolazione e delle popolazioni con cui interagisce. Questo fatto trasforma il problema dell'evoluzione da un problema di ottimizzazione a un problema di teoria dei giochi — con tutte le impossibilità che la teoria dei giochi comporta.

Il sistema coevolutivo più studiato è quello preda-predatore. Sia $x = (x_1, \dots, x_n)$ il vettore delle frequenze dei genotipi della preda e $y = (y_1, \dots, y_m)$ il vettore delle frequenze dei genotipi del predatore. L'idoneità del genotipo i della preda dipende da y (quanto efficacemente i diversi genotipi del predatore cacciano il genotipo i), e l'idoneità del genotipo j del predatore dipende da x (quanto abbondante e facile da catturare è ciascun genotipo della preda).

$$W_i^{\text{preda}}(y) = \sum_j y_j A_{ij} \quad W_j^{\text{pred}}(x) = \sum_i x_i B_{ij}$$

dove A_i è la riduzione di idoneità della preda i dovuta al predatore j , e B_j è il guadagno di idoneità del predatore j per la cattura della preda i . In questo sistema di equazioni accoppiate, l'evoluzione di ciascuna specie cambia il paesaggio adattativo dell'altra.

La dinamica coevolutiva nella formulazione più semplice (due alleli per specie, con matrice di interazione che favorisce la "corrispondenza" tra preda e predatore — come nelle interazioni pianta-patogeno dove il patogeno infetta la pianta soltanto se ha il "gene di virulenza" corrispondente al "gene di resistenza" della pianta) produce oscillazioni cicliche nelle frequenze alleliche:

$$dx_1/dt = x_1(1-x_1)(Ay_1 - A(1-y_1)) \quad dy_1/dt = y_1(1-y_1)(Bx_1 - B(1-x_1))$$

dove x_1 e y_1 sono le frequenze degli alleli "attivi" in preda e predatore, e A e B sono le intensità delle interazioni. Il sistema non ha un punto fisso stabile interno: le traiettorie sono chiuse (oscillazioni conservative) per la simmetria del sistema, e in presenza di stocasticità derivano lentamente verso la frontiera (fissazione di un allele) per effetto della deriva genetica.

VII.2 L'ipotesi della Regina Rossa

L'ipotesi della "Red Queen" (Van Valen, 1973) propone che le specie debbano evolvere continuamente per mantenere la propria idoneità in un ambiente bioticamente mutevole — come la Regina Rossa nel paese delle meraviglie di Lewis Carroll che deve correre più veloce che può per rimanere nello stesso posto. In un ambiente in cui i parassiti e i patogeni evolvono per sfruttare le caratteristiche genetiche degli ospiti, un ospite che non evolve viene progressivamente dominato da parassiti specializzati — ma l'ospite che evolve crea a sua volta nuove pressioni selettive sui parassiti.

La formalizzazione dell'ipotesi della Red Queen può essere espressa in termini di un gioco coevolutivo a somma zero. Siano $W_i(y)$ l'idoneità del genotipo i dell'ospite e $V(x)$ l'idoneità del genotipo j del parassita. Per un'interazione "gene-per-gene" (il parassita con il gene di virulenza j infetta soltanto l'ospite con il gene di resistenza i):

$$W_i(y) = 1 - c_h \cdot \sum_j y_j \delta_{ij} \quad V(x) = \sum_i x_i \delta_{ij} - c_p$$

dove δ_{ij} è 1 se il parassita j può infettare l'ospite i e 0 altrimenti, c_h è il costo dell'allele di resistenza per l'ospite, e c_p è il costo del gene di virulenza per il parassita. Il gioco ha la struttura di un gioco a somma zero — ogni guadagno del parassita è una perdita dell'ospite e viceversa — e in questa classe di giochi il teorema minimax di Von Neumann

garantisce l'esistenza di un equilibrio di Nash misto ma non garantisce la convergenza a quell'equilibrio attraverso dinamiche evolutive.

Come mostrato per il caso RPS (Sasso-Carta-Forbice) nella Monografia XXII del ciclo precedente, le dinamiche evolutive nei giochi a somma zero producono oscillazioni cicliche invece di convergere all'equilibrio: la frequenza dell'allele di resistenza dell'ospite sale quando è raro, scende quando è comune; la frequenza del gene di virulenza del parassita la insegue con un ritardo di fase. Il sistema coevolve senza mai raggiungere l'equilibrio.

VIII. Il principio di parsimonia evolutiva e i suoi limiti

VIII.1 La parsimonia filogenetica come inferenza bayesiana

La ricostruzione filogenetica — la determinazione dell'albero della vita dai dati molecolari — si basa sul principio di parsimonia (Camin e Sokal, 1965): tra tutti gli alberi filogenetici compatibili con i dati, si preferisce quello che richiede il minor numero di eventi evolutivi (mutazioni, perdite di caratteri, inversioni genomiche). Questo principio è formalmente una versione del rasoio di Occam computazionale: la spiegazione più semplice compatibile con i dati è la più probabile.

La giustificazione bayesiana della parsimonia è la seguente. Sia T un albero filogenetico e D i dati molecolari osservati. La probabilità a posteriori dell'albero è

$$P(T \mid D) \propto P(D \mid T) \cdot P(T)$$

Per un modello di sostituzione molecolare stazionario con tasso di mutazione piccolo μ , la verosimiglianza $P(D \mid T)$ è massima per l'albero che richiede il minor numero di mutazioni — l'albero più parsimonioso. In questo limite, la parsimonia e la massima verosimiglianza coincidono (Felsenstein, 1981).

Tuttavia, la parsimonia fallisce sistematicamente quando i tassi di evoluzione variano tra i rami dell'albero — il cosiddetto "long-branch attraction" (LBA). Quando due rami hanno tassi di mutazione molto alti rispetto agli altri, i caratteri che si trovano su quei rami tendono a somigliarsi per convergenza casuale (omoplasia) piuttosto che per discendenza comune. La parsimonia interpreta questa somiglianza come omologia (carattere condiviso per discendenza) e raggruppa

erroneamente i due rami ad alta evoluzione, anche se non sono filogeneticamente affini.

Il limite formale della parsimonia come stimatore filogenetico è un caso particolare di inconsistenza statistica: esiste un regime di parametri (alti tassi di evoluzione differenziali) in cui la parsimonia converge a una stima sbagliata man mano che cresce la quantità di dati. Più dati non aiutano — peggiorano la stima. Questa inconsistenza è una forma di impossibilità strutturale: nessun metodo di stima può essere simultaneamente consistente, efficiente e computazionalmente trattabile per tutti i possibili alberi filogenetici e tutti i possibili modelli di evoluzione.

VIII.2 La "vincitrice è già decisa" e l'assenza di teleologia

Il tema unificante dei limiti dell'ottimizzazione evolutiva è l'assenza di teleologia: l'evoluzione non "sa" dove sta andando e non può anticipare i cambiamenti ambientali futuri. La selezione naturale è un algoritmo miope — ottimizza l'idoneità nell'ambiente corrente senza considerazione per l'ambiente futuro.

Questa miopia ha conseguenze strutturali. Primo, i caratteri che oggi sembrano maladattativi possono essere esito di pressioni selettive passate che non esistono più: molti caratteri umani (preferenza per i cibi grassi e dolci, riflessi di paura per serpenti e ragni, tendenza all'etnocentrismo) erano adattativi nell'ambiente evolutivo di adattamento (EEA) del Pleistocene ma sono disadattativi nella modernità. Secondo, l'evoluzione può rimanere intrappolata in massimi locali del paesaggio adattativo corrente che si rivelano subottimali in futuro — la cosiddetta "contingenza evolutiva".

La contingenza evolutiva è stata formalizzata attraverso gli esperimenti di evoluzione a lungo termine di Richard Lenski (1991 e seguiti): popolazioni di *Escherichia coli* mantenute per più di 70.000 generazioni in condizioni costanti hanno seguito traiettorie evolutive diverse in repliche parallele, pur partendo dallo stesso genotipo e vivendo negli stessi ambienti. Alcune linee hanno acquisito la capacità di metabolizzare il citrato come fonte di carbonio — un'innovazione rara che non si è verificata in tutte le repliche nonostante le medesime pressioni selettive. La contingenza storica — la specificità delle mutazioni casuali che si sono verificate in ciascuna linea nelle prime migliaia di generazioni — ha determinato quale traiettoria evolutiva è stata percorsa e se l'innovazione si è verificata o no.

IX. Conclusioni: la selezione come confine tra possibile e necessario

Il percorso di questa monografia ha illustrato perché l'evoluzione per selezione naturale, pur essendo il meccanismo più potente noto per la generazione di complessità adattiva, opera entro vincoli strutturali che la rendono intrinsecamente limitata come algoritmo di ottimizzazione.

Il teorema fondamentale di Fisher garantisce soltanto che la selezione aumenta l'idoneità media al tasso della varianza additiva — una garanzia locale, transitoria e relativa alla componente selettiva del cambiamento evolutivo, che viene tipicamente più che controbilanciata dai cambiamenti ambientali e co-evolutivi. La deriva genetica introduce una componente stocastica irriducibile che può fissare mutazioni deleterie e perdere mutazioni vantaggiose, particolarmente nelle popolazioni piccole. Il modello NK mostra che per paesaggi adattativi sufficientemente accidentati (alta epistasi), il hill-climbing si arena in minimi locali con alta probabilità, e la probabilità di trovare il massimo globale è esponenzialmente piccola. Il teorema No Free Lunch afferma che la selezione naturale non è migliore di un algoritmo casuale quando mediata sulla distribuzione uniforme di tutti i paesaggi adattativi possibili: la sua efficacia dipende interamente dalla struttura correlata dei paesaggi biologici reali, struttura che non è garantita e che si dissolve in certi regimi. Il ratchet di Müller mostra che le popolazioni asessuate accumulano mutazioni deleterie irreversibilmente, un problema che il sesso risolve attraverso una ricombinazione che è essa stessa un meccanismo con propri costi e vincoli.

Questi limiti non diminuiscono la grandiosità del risultato evolutivo — la biodiversità terrestre è il prodotto di quattro miliardi di anni di evoluzione, e la complessità dei sistemi biologici supera di molti ordini di grandezza quella di qualsiasi sistema progettato dall'ingegneria. Ma comprendere i limiti è il prerequisito per comprendere la struttura del risultato: il cervello umano è la risposta evolutiva ai problemi di rappresentazione analizzati nella Monografia I, realizzata entro i vincoli di ottimizzazione analizzati qui. Nelle monografie successive si vedrà come l'evoluzione abbia costruito — attraverso miliardi di tentativi e dentro questi vincoli precisi — soluzioni sorprendentemente efficaci a problemi che la teoria formale dichiara insolubili in linea generale.

Monografia III

Il problema del frame nell'ecologia cognitiva

Affordance, percezione diretta, inferenza bayesiana vincolata e i limiti strutturali della rilevanza

I. Il problema della rilevanza come impossibilità computazionale

Tra tutti i problemi che un sistema cognitivo deve risolvere per sopravvivere in un ambiente fisico complesso, il più silenzioso e il più devastante è il seguente: in ogni momento, l'ambiente offre all'agente una quantità potenzialmente illimitata di informazione sensoriale, e soltanto una piccola frazione di essa è rilevante per le decisioni immediate. Il problema non è la mancanza di informazione — è la sovrabbondanza. Il sistema deve scegliere a cosa prestare attenzione, quali aspetti dell'ambiente integrare nella propria rappresentazione corrente, e quali ignorare senza per questo perdere informazioni critiche per la sopravvivenza.

Questo problema — noto in filosofia come "problema del frame" dopo che McCarthy e Hayes (1969) lo formularono nel contesto dell'intelligenza artificiale simbolica — ha una struttura che si rivela, all'analisi formale, non risolvibile in linea generale da nessun algoritmo finito. La sua manifestazione biologica, ossia il modo in cui i sistemi cognitivi reali gestiscono questa impossibilità senza collassare, è il tema centrale di questa monografia.

La versione originale di McCarthy e Hayes è la seguente. In un sistema di pianificazione basato sulla logica classica, un agente deve ragionare sugli effetti delle proprie azioni sullo stato del mondo. Il problema del frame è che la maggior parte delle proprietà del mondo rimane invariata dopo un'azione — se l'agente sposta una scatola, il colore del cielo rimane lo stesso, la temperatura dell'acqua rimane la stessa, il numero degli atomi nell'universo rimane lo stesso. Ma in un sistema logico, ogni proprietà che non è esplicitamente dedotta come cambiata deve essere assunta come invariata — il che richiede una quantità potenzialmente infinita di assiomi del tipo "l'azione A non cambia la proprietà P", uno per ogni coppia (azione, proprietà) con effetto nullo.

Il problema del frame nell'ecologia cognitiva è la versione biologica di questo problema logico: come un sistema cognitivo con risorse computazionali finite determina, per ogni situazione ambientale, quali

proprietà dell'ambiente sono rilevanti per le proprie decisioni correnti, senza dover valutare esplicitamente ogni possibile proprietà. La risposta biologica non è un algoritmo generale — non esiste — ma è un insieme di soluzioni ecologicamente vincolate che funzionano nei domini evolutivi in cui il sistema è stato selezionato, e falliscono sistematicamente fuori da quegli stessi domini.

II. La teoria delle affordance di Gibson: percezione come rilevazione di possibilità d'azione

II.1 Il concetto di affordance e la sua struttura relazionale

James Jerome Gibson (1966, 1979) propose una teoria della percezione radicalmente diversa sia dall'approccio computazionale-rappresentazionale sia dall'approccio behaviorista. Per Gibson, la percezione non è la costruzione di una rappresentazione interna del mondo a partire dal flusso sensoriale — è la rilevazione diretta di proprietà del mondo che sono immediatamente rilevanti per l'azione dell'organismo. Il concetto centrale è quello di affordance: una proprietà dell'ambiente che offre ("affords") una possibilità d'azione a un particolare organismo.

La definizione formale di Gibson è intenzionalmente relazionale: un'affordance è una proprietà dell'ambiente che è specificata rispetto alle capacità dell'agente. Una superficie piatta e resistente afforda il cammino per un umano ma non per un pesce; una bacca afforda il mangiare per un uccello con il becco adatto ma non per uno con un becco diverso; un ramo afforda l'appiglio per una scimmia ma non per un bovino. L'affordance non è né una proprietà puramente soggettiva (nell'agente) né puramente oggettiva (nell'ambiente): è una proprietà della relazione tra agente e ambiente.

La struttura matematica più appropriata per formalizzare le affordance è quella di una relazione binaria

$$\text{Aff} \subseteq E \times O$$

dove E è lo spazio delle proprietà ambientali e O è lo spazio degli organismi (caratterizzati dalle loro capacità motorie, percettive e metaboliche). Un'affordance è una coppia $(e, o) \in \text{Aff}$ che dice: la proprietà ambientale e è rilevante — offre una possibilità d'azione —

per l'organismo o. La percezione ecologica è la proiezione di questa relazione sull'asse dell'organismo:

$$\text{Aff}_o = \{e \in E : (e, o) \in \text{Aff}\}$$

ovvero l'insieme delle proprietà ambientali che sono rilevanti per un dato organismo. Il problema del frame biologico è computare efficientemente Aff_o in tempo reale, senza dover valutare esplicitamente ogni $e \in E$.

II.2 L'informazione ottica di flusso e la struttura invariante

Il merito principale della teoria gibsoniana non è il concetto di affordance in sé — che in questa forma è quasi tautologico — ma l'indicazione di come gli organismi rilevino le affordances senza calcolarle: attraverso la struttura dell'informazione ambientale che è già organizzata in modo da specificare le proprietà rilevanti, senza richiedere un'inferenza computazionale intermedia.

Il caso più elaborato da Gibson è la percezione visiva del moto. Quando un osservatore si muove nell'ambiente o quando oggetti si muovono nell'ambiente, la luce riflessa dagli oggetti produce un pattern di flusso ottico sulla retina. Gibson mostra che questo flusso ottico contiene informazione strutturale su proprietà ecologicamente rilevanti dell'ambiente — profondità, struttura delle superfici, moto degli oggetti — in una forma che non richiede calcoli di triangolazione o di stereopsi.

Il campo di flusso ottico di un osservatore che si muove con velocità v in un ambiente statico è il campo vettoriale sulla sfera di proiezione retinica:

$$F(\theta, \varphi) = -v_{\text{retina}}(\theta, \varphi) / d(\theta, \varphi)$$

dove θ e φ sono le coordinate sferiche sulla retina, $d(\theta, \varphi)$ è la distanza dell'oggetto nella direzione (θ, φ) dall'osservatore, e v_{retina} è la componente della velocità dell'osservatore nel piano retinico perpendicolare alla direzione di sguardo. Il campo di flusso ha un singolo punto di espansione (il focus of expansion, FOE) nella direzione del moto dell'osservatore:

$$\text{FOE} = \text{direzione di } v$$

La posizione del FOE specifica direttamente la direzione di moto dell'osservatore — senza triangolazione, senza stereopsi, senza modello esplicito dell'ambiente. La velocità di espansione intorno al FOE

specifica il tempo al contatto (time to contact, TTC) con le superfici nella direzione di moto:

$$TTC = d / |v| = \theta_{size} / (d\theta_{size}/dt)$$

dove θ_{size} è la dimensione angolare dell'oggetto e $d\theta_{size}/dt$ è la sua velocità di espansione angolare. Il rapporto $\theta_{size} / (d\theta_{size}/dt)$ — chiamato tau (τ) da Lee (1976) — è un invariante ottico che specifica il tempo al contatto indipendentemente dalla distanza assoluta o dalla velocità assoluta, e senza richiedere alcuna conoscenza esplicita di queste grandezze.

II.3 La critica computazionale e i limiti della percezione diretta

La teoria gibsoniana ha subito una critica formale sistematica da parte della scuola computazionale — David Marr, Tomaso Poggio, Shimon Ullman — che ne ha accettato il programma descrittivo (identificare le invarianti ottiche) ma ne ha rifiutato il programma esplicativo (queste invarianti vengono percepite "direttamente", senza elaborazione computazionale interna).

La critica principale è la seguente. Anche ammettendo che l'informazione ambientale sia strutturata in modo da contenere le invarianti ecologiche, la rilevazione di queste invarianti da parte del sistema nervoso richiede computazioni non banali. Il flusso ottico, ad esempio, richiede per la sua estrazione dal segnale retinico la stima delle derivate spazio-temporali del segnale di luminanza — un'operazione che è soggetta a rumore, a ambiguità di apertura (il fenomeno per cui il moto di un contorno è misurabile solo nella componente perpendicolare al contorno, non nella componente parallela) e a problemi di segmentazione figura-sfondo.

La formula per l'estrazione del flusso ottico come soluzione dell'equazione di vincolo sul flusso ottico (Horn e Schunck, 1981) è:

$$(\partial I / \partial x)u + (\partial I / \partial y)v + \partial I / \partial t = 0$$

dove $I(x,y,t)$ è la luminanza al punto (x,y) al tempo t , e $(u,v) = (dx/dt, dy/dt)$ è il flusso ottico locale. Questa equazione ha due incognite (u e v) per ogni punto e soltanto un'equazione: il sistema è sotto-determinato, il che significa che il flusso ottico locale non è determinato unicamente dal segnale di luminanza. La soluzione richiede l'imposizione di vincoli aggiuntivi (regolarità del campo di flusso, preservazione dello spigolo, modello del moto rigido) che sono ipotesi sul mondo — non caratteristiche del segnale sensoriale.

L'ambiguità di apertura è particolarmente istruttiva. Un recettore di moto (un neurone sensibile alla direzione del moto) che risponde a una finestra locale del campo visivo non può distinguere diversi vettori di flusso che proiettano la stessa componente perpendicolare al bordo. La percezione del moto globale richiede l'integrazione delle risposte di molti recettori locali — un processo computazionale non banale che il cervello risolve attraverso meccanismi specifici (integrazione nel MT/V5 per il moto coerente globale) che presuppongono ipotesi sul tipo di moto (rigido vs. non-rigido, piano vs. 3D).

III. Il problema del frame come problema di rilevanza: struttura formale

III.1 La formulazione logica del problema del frame

McCarthy e Hayes (1969) formularono il problema del frame nella logica situazionale come segue. Sia S lo spazio delle situazioni (stati del mondo), A lo spazio delle azioni, e $R \subseteq S \times A \times S$ la relazione di transizione ($R(s, a, s')$ significa che eseguendo l'azione a nello stato s si raggiunge lo stato s'). Per ogni coppia (a, s) , le proprietà del mondo che non sono direttamente affette dall'azione a devono essere esplicitamente elencate come invariate: questa è la "specificazione delle conseguenze indirette" o frame problem.

In un mondo con n proprietà booleane e m azioni, il numero di assiomi del frame necessari è potenzialmente $n \times m$ — uno per ogni coppia (proprietà, azione) per cui la proprietà rimane invariata dopo l'azione. Per $n = 10^6$ proprietà e $m = 10^3$ azioni possibili, questo richiede 10^9 assiomi del frame — un numero intrattabile sia in termini di specificazione (nessun programmatore può enumerarli tutti) sia in termini di elaborazione (nessun sistema logico può inferire su 10^9 assiomi in tempo reale).

Le soluzioni proposte all'interno del paradigma logico — la Circumscription di McCarthy (1980), la Default Logic di Reiter (1980), il ragionamento non monotono in generale — cercano di minimizzare il numero di cambiamenti tra stati o di specificare soltanto le conseguenze dirette delle azioni (le "leggi di effetto") lasciando implicitamente invariato tutto il resto. Ma queste soluzioni hanno un prezzo formale: sono computazionalmente intractabili nel caso generale (il ragionamento non monotono è Π_2^P -completo nella gerarchia polinomiale).

Teorema (intrattabilità del frame, Liberatore e Schaerf, 1998). Il problema di decidere se una formula proposizionale è una conseguenza della Circumscription di una base di conoscenza è Π_2^P -completo.

La Π_2^P -completezza significa che il problema è intrinsecamente più difficile dei problemi NP-completi: anche se $P = NP$ (il che renderebbe i problemi NP risolvibili in tempo polinomiale), i problemi Π_2^P rimarrebbero intrattabili. Nessun computer con risorse polinomiali può risolvere il problema del frame nel caso generale.

III.2 Il problema del frame nel ragionamento quotidiano

La versione biologica del problema del frame si pone in ogni situazione in cui un organismo deve aggiornare la propria rappresentazione del mondo dopo un'azione o un evento. Quando un umano sposta un oggetto dal tavolo alla sedia, deve aggiornare la rappresentazione della posizione dell'oggetto (cambiata), mantenere invariata la rappresentazione del colore dell'oggetto (non cambiata), aggiornare o meno la rappresentazione della posizione degli altri oggetti sul tavolo (non spostati), e così via per ogni proprietà del mondo rappresentata.

Il fatto che questo aggiornamento avvenga automaticamente e senza sforzo evidente nell'esperienza quotidiana ha indotto a pensare che il problema del frame non sia un problema reale per i sistemi cognitivi biologici. Questa impressione è ingannevole: il problema è risolto non computando la risposta (che sarebbe intrattabile) ma aggirando la necessità della computazione attraverso architetture percettive che selezionano l'informazione rilevante prima che essa entri nel sistema di ragionamento esplicito.

La struttura di questo aggiramento è la seguente. Il sistema sensoriale-percettivo del cervello è organizzato in modo da estrarre specifiche invarianti dall'informazione ambientale — quelle invarianti che sono state biologicamente selezionate perché segnalano eventi e proprietà rilevanti per la sopravvivenza. Queste invarianti sono disponibili al sistema cognitivo già preelaborate, senza che il sistema debba valutare esplicitamente ogni possibile proprietà dell'ambiente.

La struttura matematica appropriata per descrivere questo meccanismo è quella dei filtri bayesiani — e specificamente del filtro di Kalman e dei suoi generalizzazioni non lineari.

III.3 Il filtro di Kalman come soluzione al frame per sistemi lineari

Considerato un agente che si muove in un ambiente in cui lo stato del mondo x_t evolve secondo una dinamica lineare con rumore gaussiano:

$$x_{t+1} = F x_t + B u_t + w_t, w_t \sim N(0, Q)$$

dove $x_t \in \mathbb{R}^n$ è il vettore di stato al tempo t , F è la matrice di transizione di stato, B è la matrice di controllo, u_t è il vettore dei comandi dell'agente, e w_t è il rumore di processo con covarianza Q . Le osservazioni y_t sono

$$y_t = H x_t + v_t, v_t \sim N(0, R)$$

dove H è la matrice di osservazione e v_t è il rumore di osservazione con covarianza R .

Il filtro di Kalman aggiorna la distribuzione posteriore $p(x_t | y_{1:t})$ in modo ottimale, in due passi.

Passo di predizione (propagazione del prior):

$$\hat{x}_{t|t-1} = F \hat{x}_{t-1|t-1} + B u_t P_{t-1|t-1} = F P_{t-1|t-1} F^T + Q$$

Passo di aggiornamento (correzione bayesiana):

$$K_t = P_{t|t-1} H^T (H P_{t|t-1} H^T + R)^{-1} \hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (y_t - H \hat{x}_{t|t-1}) P_{t|t} = (I - K_t H) P_{t|t-1}$$

dove K_t è il guadagno di Kalman — il peso con cui la nuova osservazione y_t aggiorna la stima dello stato rispetto alla predizione.

Il filtro di Kalman risolve il problema del frame per sistemi lineari gaussiani in modo esatto e computazionalmente efficiente: il costo per ogni passo di aggiornamento è $O(n^3)$ (dove n è la dimensione dello stato), indipendentemente dal numero di osservazioni precedenti. La soluzione al problema del frame è implicita nella struttura del filtro: non è necessario specificare esplicitamente quali proprietà del mondo rimangono invariate dopo un'azione — questo è codificato implicitamente nella matrice di transizione F , che specifica come ciascuna componente dello stato evolve in funzione delle altre.

Tuttavia, il filtro di Kalman risolve il problema del frame soltanto per i sistemi per cui lo stato rilevante è di dimensione fissa e note a priori. In ambienti reali, la dimensione dello stato rilevante è variabile: quando apro una porta, il numero di oggetti nel mio campo d'attenzione cambia; quando entro in una stanza nuova, devo aggiornare la mia rappresentazione con oggetti che non erano nel mio modello del

mondo. Queste variazioni della dimensione dello stato sono il cuore del problema del frame biologico.

IV. L'ecologia bayesiana: inferenza ottimale vincolata

IV.1 Il cervello come macchina bayesiana

La formulazione bayesiana della percezione — il cervello come un sistema che computa la distribuzione posteriore sullo stato del mondo condizionatamente alle osservazioni sensoriali — è diventata il paradigma dominante nella neuroscienza computazionale degli ultimi trent'anni. La formulazione è dovuta a vari autori ma ha trovato la sua espressione più sistematica nel lavoro di Knill e Whitman Richards (1996), Ernst e Banks (2002), e Körding e Wolpert (2004).

L'idea fondamentale è che la percezione non è la lettura passiva di dati sensoriali, bensì un processo inferenziale attivo in cui le informazioni sensoriali ambigue vengono integrate con le conoscenze a priori del sistema (codificate nelle distribuzioni prior) per produrre la stima più probabile dello stato del mondo. La formula di Bayes applicata alla percezione è:

$$p(s | r) \propto p(r | s) \cdot p(s)$$

dove $s \in S$ è lo stato del mondo (ad esempio la posizione di un oggetto, l'orientamento di una superficie, la causa di un odore), $r \in R$ è la risposta sensoriale (la luminanza retinica, il pattern di pressione sulla coclea, la concentrazione di un odorifero), $p(r | s)$ è la verosimiglianza della risposta sensoriale dato lo stato del mondo (il modello sensoriale), e $p(s)$ è il prior sulla distribuzione degli stati del mondo.

La percezione bayesiana ottimale in questo schema corrisponde alla scelta di

$$\hat{s} = \operatorname{argmax}_s p(s | r) = \operatorname{argmax}_s [p(r | s) \cdot p(s)]$$

per la stima MAP (Maximum A Posteriori) o la media della distribuzione posteriore per la stima che minimizza l'errore quadratico medio.

IV.2 La risoluzione delle ambiguità percettive come inferenza bayesiana

L'evidenza sperimentale per il modello bayesiano della percezione è vasta. Ernst e Banks (2002) studiarono l'integrazione visivo-tattile nella percezione della dimensione degli oggetti. I soggetti dovevano stimare l'altezza di una barra che vedevano e toccavano simultaneamente, in condizioni in cui l'informazione visiva e quella tattile erano deliberatamente messe in conflitto attraverso un prisma che distorceva la percezione visiva delle dimensioni.

Il modello bayesiano ottimale per l'integrazione di due sorgenti di informazione con varianze σ_v^2 (visiva) e σ_t^2 (tattile) produce la stima

$$\hat{s} = (w_v s_v + w_t s_t) / (w_v + w_t)$$

con pesi

$$w_v = 1/\sigma_v^2, w_t = 1/\sigma_t^2$$

e varianza della stima integrata

$$\sigma_{int}^2 = (\sigma_v^2 \cdot \sigma_t^2) / (\sigma_v^2 + \sigma_t^2) < \min(\sigma_v^2, \sigma_t^2)$$

L'integrazione ottimale riduce la varianza della stima al di sotto della varianza di ciascuna sorgente individuale. Ernst e Banks trovarono che i soggetti si comportano in accordo con questa previsione: i pesi assegnati a ciascuna sorgente sono proporzionali all'inverso della varianza di quella sorgente, e la varianza della stima integrata è esattamente quella prevista dall'integrazione ottimale.

Analoghi risultati furono ottenuti per l'integrazione visivo-vestibolare nella percezione del moto (Angelaki et al., 2009), per l'integrazione audio-visiva nella percezione del linguaggio (effetto McGurk: quando il suono "ba" è sincronizzato con il labiale "ga", si percepisce "da"), e per l'integrazione propriocettiva-visiva nella percezione della posizione degli arti.

IV.3 Il prior come conoscenza ecologica: la soluzione biologica al problema del frame

La soluzione biologica al problema del frame è codificata nella struttura dei prior bayesiani — nelle distribuzioni a priori che il sistema percettivo utilizza per interpretare i segnali sensoriali ambigui. Questi prior non sono stati "calcolati" esplicitamente dal sistema cognitivo nell'arco della vita dell'individuo: sono stati selezionati evolutivamente nell'arco delle generazioni, sulla base della distribuzione statistica delle proprietà ecologicamente rilevanti nell'ambiente evolutivo di adattamento.

Il prior più studiato è il prior di "luce dall'alto": il sistema visivo assume che le sorgenti luminose siano collocate in posizione superiore rispetto agli oggetti, perché nell'ambiente naturale la luce solare proviene sempre dall'alto. Quando un'immagine bidimensionale di una superficie con ombre viene capovolta, la percezione della rilievo si inverte — ciò che sembrava convesso sembra concavo e viceversa — perché il prior di luce-dall'alto interpreta l'orientamento dell'ombra come indicatore della direzione della curvatura.

Il prior sulla rigidità degli oggetti fisici è un altro esempio. Quando un insieme di punti luminosi si muove in modo coerente con la proiezione prospettica di un oggetto rigido tridimensionale, il sistema visivo percepisce un oggetto rigido tridimensionale anche quando la struttura tridimensionale è irrecuperabile dai dati (l'informazione proiettiva bidimensionale è compatibile con infinite strutture tridimensionali). Ullman (1979) dimostrò che tre punti in quattro frame sono sufficienti per la percezione di struttura tridimensionale rigida (theorem of structure from motion), e questo avviene perché il sistema presuppone la rigidità.

La struttura formale del prior di rigidità è un vincolo sullo spazio degli stati rilevanti:

$$S_{\text{rigido}} = \{s \in S_{\text{3D}} : \exists R \in SO(3), T \in \mathbb{R}^3 : s_t = R_t s_0 + T_t \forall t\}$$

Il prior di rigidità assegna probabilità zero a tutti gli stati non rigidi — la dimensione rilevante dello spazio degli stati è ridotta da $|S_{\text{3D}}|$ (potenzialmente infinita) a $|S_{\text{rigido}}|$ (parametrizzabile con 6 gradi di libertà: rotazione $SO(3)$ e traslazione $\mathbb{R}^3$). Questa riduzione della dimensionalità dello spazio degli stati è il meccanismo con cui i prior ecologici risolvono praticamente il problema del frame: restringendo lo spazio delle ipotesi rilevanti, rendono trattabile l'inferenza che altrimenti sarebbe computazionalmente impossibile.

V. I limiti ecologici dei prior: la cognizione fuori dal dominio evolutivo

V.1 I prior come soluzioni locali a problemi globali

La soluzione biologica al problema del frame — codificare i prior ecologici nell'architettura del sistema percettivo — è una soluzione locale: funziona nei domini evolutivi per cui è stata selezionata e fallisce

sistematicamente fuori da quei domini. Questo è il principio che questa sezione formalizza.

Sia P_{eco} il prior ecologico codificato nel sistema percettivo di un organismo — la distribuzione di probabilità sulle proprietà del mondo che è stata selezionata evolutivamente sulla base delle statistiche ambientali nell'EEA (Evolutionary Environment of Adaptedness). Sia P_{dom} la distribuzione delle proprietà del mondo nel dominio di deployment corrente dell'organismo — che può coincidere o meno con l'EEA.

La performance del sistema percettivo nel dominio dom è determinata dalla distanza tra P_{eco} e P_{dom} :

$$Performance(dom) = f(D(P_{eco}, P_{dom}))$$

dove D è una misura di discrepanza tra distribuzioni (ad esempio la divergenza di Kullback-Leibler $D_{KL}(P_{dom} || P_{eco})$) e f è una funzione decrescente. Quando $dom = EEA$, la distanza è zero e la performance è ottimale; quando $dom \neq EEA$, la distanza può essere arbitrariamente grande e la performance può essere arbitrariamente scarsa.

Esempi concreti di fallimento del prior ecologico fuori dall'EEA:

Il prior di luce-dall'alto è calibrato sull'illuminazione solare, che è sempre superiore. In ambienti con illuminazione artificiale laterale o inferiore (tunnel, caverne, stanze con finestre basse), la percezione della rilievo è sistematicamente distorta.

Il prior di rigidità funziona per oggetti solidi naturali (rocce, alberi, animali) ma fallisce per liquidi e plasmi, i cui moti non sono rigidi, portando a percezioni erronee della struttura tridimensionale di fiamme, nebbie e acqua.

Il prior sulle scale di distanza è calibrato sull'ambiente aperto del savana africano (dove la maggior parte dell'evoluzione umana è avvenuta): in ambienti urbani, con edifici ad alta densità e prospettive compresse, i prior di distanza producono sistematicamente stime errate.

V.2 La formalizzazione attraverso la teoria delle decisioni ecologiche

Gigerenzen e Todd (1999) hanno proposto che i sistemi cognitivi umani implementino "euristiche veloci e frugali" — procedure decisionali che usano poca informazione e pochi calcoli, ma che si adattano alla struttura statistica degli ambienti per cui sono state selezionate. Queste

euristiche sono la versione cognitiva dei prior ecologici bayesiani: non computano esplicitamente la distribuzione posteriore, ma implementano approssimazioni che convergono alla risposta bayesiana ottimale nell'ambiente per cui sono state selezionate.

L'euristica "Take the Best" è un esempio paradigmatico. Quando si deve scegliere quale dei due oggetti ha un certo valore (ad esempio, quale delle due città ha la popolazione maggiore), l'euristica cerca la prima cue (il primo indicatore disponibile) che discrimina tra i due oggetti e usa soltanto quella cue per la decisione, ignorando tutte le altre.

Formalmente, siano $c_1, c_2, \dots, c$ le cue disponibili, ordinate per validità decrescente ($v_1 \geq v_2 \geq \dots \geq v$ dove v_i è la correlazione tra la cue i e il criterio). L'euristica Take the Best:

- (1) Cerca la cue con validità più alta che discrimina tra i due oggetti (cue con valori diversi).
- (2) Prende la decisione sulla base di quella sola cue.
- (3) Ignora tutte le cue successive.

La performance dell'euristica rispetto alla regressione logistica multipla (che usa tutte le cue in modo ottimale) dipende dalla struttura ambientale. In ambienti con struttura "a J" delle validità delle cue (dove la prima cue è molto più valida delle successive, come in molti ambienti naturali), Take the Best è quasi ottimale e spesso supera la regressione multipla in termini di generalizzazione a dati nuovi. In ambienti con struttura "piatta" delle validità (dove tutte le cue hanno validità simile), Take the Best è sub-ottimale rispetto alla regressione multipla.

Il risultato formale di Martignon e Hoffrage (2002) è:

Teorema (ottimalità ecologica di Take the Best). In un ambiente con struttura di correlazioni "a J" — dove la prima cue ha correlazione con il criterio $c_1 \geq c_2 + \dots + c$ — l'euristica Take the Best produce la stessa risposta della regressione logistica multipla con probabilità

$$P(\text{accordo}) \geq 1 - \epsilon$$

per ϵ piccolo che dipende dalla curvatura del profilo delle validità.

Questa è la formalizzazione quantitativa del principio "less is more" gibsoniano: in ambienti con la struttura appropriata, meno informazione — usata in modo appropriato — produce decisioni non peggiori di più informazione usata in modo più complesso. La "più appropriato" è la chiave: l'euristica è ottimale soltanto nell'ambiente per cui è adattata.

VI. Il problema del frame per eventi sociali: la cognizione sociale e i suoi vincoli

VI.1 La cognizione sociale come rilevamento di intenzioni

Nelle interazioni sociali, il problema del frame assume una forma specialmente acuta: quale aspetto del comportamento di un altro agente è rilevante per interpretare le sue intenzioni e prevedere le sue azioni future? In un ambiente fisico, il numero di proprietà rilevanti è grande ma finito e relativamente stabile; in un ambiente sociale, le intenzioni degli altri agenti sono nascoste, mutevoli e legate in modo non lineare al comportamento osservabile.

La soluzione evolutiva umana al problema del frame sociale è la "Teoria della Mente" (Theory of Mind, ToM) — la capacità di attribuire stati mentali (credenze, desideri, intenzioni) ad altri agenti e di usare queste attribuzioni per prevederne il comportamento. La ToM è una soluzione al problema del frame sociale perché restringe drasticamente lo spazio delle ipotesi rilevanti: invece di considerare ogni possibile stato futuro dell'altro agente (uno spazio potenzialmente infinito), il sistema cognitivo considera soltanto gli stati futuri coerenti con le intenzioni attribuite all'agente — uno spazio molto più ristretto.

La ToM matura negli esseri umani intorno ai 4 anni attraverso la comprensione della "falsa credenza" (false belief): la capacità di rappresentare che un altro agente possa avere credenze che sono false (non corrispondenti alla realtà come la conosce l'osservatore). Il classico test di Wimmer e Perner (1983) — la storia di Maxi che ripone il cioccolato in un cassetto e se ne va, dopo di che la madre lo sposta in un altro cassetto — richiede che il bambino preveda dove Maxi cercherà il cioccolato al suo ritorno (nel posto in cui l'ha riposto, non nel posto in cui si trova ora), dimostrando la comprensione che Maxi ha credenze che differiscono dalla realtà corrente.

VI.2 Il sistema a due processi e la sua struttura

La soluzione cognitiva al problema del frame non è unica: i sistemi cognitivi biologici implementano almeno due sistemi di elaborazione con proprietà diverse, che la psicologia cognitiva moderna descrive attraverso la teoria dei processi duali (Kahneman, 2011; Evans, 2008).

Il Sistema 1 (S1) è automatico, veloce, parallelo, non consapevole e guidato dai prior ecologici. Risolve il problema del frame attraverso il

riconoscimento di pattern: situazioni familiari attivano rapidamente risposte pre-calcolate senza la necessità di un ragionamento esplicito sulle proprietà rilevanti. Il Sistema 2 (S2) è controllato, lento, seriale, consapevole e capace di ragionamento esplicito. Affronta il problema del frame attraverso l'attenzione selettiva: identifica deliberatamente le proprietà rilevanti per la situazione corrente e costruisce rappresentazioni esplicite per il ragionamento.

La formalizzazione più precisa dei due sistemi usa il linguaggio della teoria del controllo ottimale gerarchico. Il S1 implementa un controllo feedforward: dato il pattern riconosciuto S, produce la risposta R(S) pre-programmata senza misurare esplicitamente le proprietà correnti dell'ambiente. Il S2 implementa un controllo feedback: misura le proprietà correnti dell'ambiente, le confronta con le aspettative, e produce correzioni quando vi è discrepanza.

Il costo computazionale del S1 è $O(1)$ per situazione familiare — il riconoscimento del pattern avviene in tempo costante una volta che il pattern è stato appreso. Il costo del S2 è $O(n \cdot m)$ dove n è il numero di proprietà valutate e m è la profondità del ragionamento — cresce con la complessità della situazione. Il vantaggio ecologico del S1 è la velocità e l'efficienza metabolica; il suo limite è la rigidità: non può adattarsi a situazioni nuove che si discostano dai pattern appresi.

VI.3 L'efficacia del S1 e i suoi fallimenti sistematici

I fallimenti del S1 nelle situazioni nuove costituiscono il catalogo classico dei bias cognitivi di Kahneman e Tversky. La struttura formale è quella di un prior mal calibrato applicato a un ambiente con distribuzione diversa dall'EEA.

La Fallacy del tasso base (base rate neglect) è l'esempio più illuminante. Il problema classico di Tversky e Kahneman (1974): "Linda ha 31 anni, è single, franca, molto intelligente. Si è laureata in filosofia. Da studentessa era molto interessata alle questioni di discriminazione e giustizia sociale e ha anche partecipato a manifestazioni antinucleari. Cosa è più probabile: (a) Linda è cassiera di banca, (b) Linda è cassiera di banca e attivista femminista?". La maggioranza dei soggetti risponde (b) — una violazione elementare della probabilità ($P(A \cap B) \leq P(A)$ sempre).

La spiegazione bayesiana è la seguente. Il prior su "tipo di persona" ha alta probabilità per "attivista femminista" dato il profilo descritto, e bassa probabilità per "cassiera di banca". Il S1 usa questo prior per valutare la probabilità degli scenari attraverso la "rappresentatività" — la somiglianza del profilo agli stereotipi delle categorie — e ignora il tasso base ($P(\text{cassiera}) \gg P(\text{cassiera e attivista})$). Il prior ecologico sul "tipo di

persona" — che nella vita quotidiana è uno strumento eccellente per prevedere il comportamento degli individui — produce inferenze scorrette quando applicato a problemi formali di probabilità, un dominio che non esiste nell'EEA.

Formalmente, il S1 implementa l'inferenza

$$P_{S1}(\text{categoria} \mid \text{descrizione}) \propto P(\text{descrizione} \mid \text{categoria}) \cdot \text{prior}_{S1}(\text{categoria})$$

dove prior_{S1} è il prior rappresentazionale (basato sulla somiglianza agli stereotipi) invece del prior statistico corretto (basato sulle frequenze di base delle categorie nella popolazione). La divergenza tra prior_{S1} e il prior statistico vero è l'origine sistematica dell'errore.

VII. La percezione dell'azione e il problema del frame inverso

VII.1 La percezione come soluzione al problema inverso

La percezione visiva, nella sua formulazione computazionale (Marr, 1982), è un problema inverso: dato il pattern di luminanza sulla retina (una proiezione bidimensionale del mondo tridimensionale), ricostruire le proprietà tridimensionali del mondo che hanno prodotto quella proiezione. Questo problema inverso è mal posto nel senso di Hadamard: non ha soluzione unica (infiniti mondi tridimensionali proiettano sulla stessa immagine bidimensionale), non è stabile (piccole variazioni dell'immagine possono corrispondere a grandi variazioni del mondo), e non dipende continuamente dai dati.

La soluzione biologica al problema inverso mal posto è la regolarizzazione: l'imposizione di vincoli a priori sullo spazio delle soluzioni che rendono il problema ben posto. I prior ecologici (rigidità, luce dall'alto, continuità delle superfici) sono forme di regolarizzazione bayesiana: selezionano tra le infinite soluzioni compatibili con i dati sensoriali quella che è più probabile alla luce della distribuzione statistica dei mondi naturali.

La struttura formale è quella dell'ottimizzazione di Tikhonov. Data la relazione tra il mondo s e l'immagine $y = Hs + \text{rumore}$ (dove H è la matrice di proiezione prospettica), la soluzione ai minimi quadrati regolarizzata è

$$\hat{s} = \operatorname{argmin}_s [\|y - Hs\|^2 + \lambda \cdot \Omega(s)]$$

dove $\Omega(s)$ è il funzionale di regolarizzazione che penalizza soluzioni "non naturali" (ad esempio la discontinuità delle superfici, la non-rigidità degli oggetti, l'illuminazione dal basso), e λ è il parametro di regolarizzazione che controlla il bilanciamento tra fedeltà ai dati e regolarità della soluzione.

Il prior bayesiano equivalente è

$$p(s) \propto \exp(-\lambda \cdot \Omega(s))$$

e la soluzione MAP è

$$\hat{s} = \operatorname{argmax}_s p(s \mid y) = \operatorname{argmax}_s [p(y \mid s) \cdot p(s)] = \operatorname{argmin}_s [-\log p(y \mid s) - \log p(s)] = \operatorname{argmin}_s [\|y - Hs\|^2 / \sigma^2 + \lambda \cdot \Omega(s)]$$

che è identicamente il problema di Tikhonov. La regolarizzazione e l'inferenza bayesiana sono due formulazioni dello stesso processo di soluzione del problema inverso con vincoli a priori.

VII.2 Il problema del frame inverso: agire per risolvere l'ambiguità percettiva

Gibson aveva osservato che la percezione non è passiva — i sistemi cognitivi mobili si muovono nell'ambiente in modo da risolvere le ambiguità percettive. Questo è il "problema del frame inverso": come l'agente decide quali azioni intraprendere per ridurre l'incertezza sulle proprietà rilevanti dell'ambiente, tra la molteplicità di azioni possibili.

La formalizzazione è quella della percezione attiva (Active Perception) di Bajcsy (1988) e dell'ottimizzazione dell'informazione nella ricerca attiva (Active Vision) di Aloimonos (1990). Un agente che ha un modello bayesiano del mondo con distribuzione posteriore $p(s \mid y_{1:t})$ può scegliere la prossima azione a_{t+1} in modo da massimizzare la riduzione attesa di incertezza:

$$a_{t+1}^* = \operatorname{argmax}_a I(S; Y_{t+1} \mid a, Y_{1:t})$$

dove $I(S; Y_{t+1} \mid a, Y_{1:t})$ è l'informazione mutua attesa tra lo stato del mondo S e l'osservazione futura Y_{t+1} se si esegue l'azione a . Massimizzare l'informazione mutua attesa corrisponde a scegliere l'azione che, in media, produrrà le osservazioni più informative sullo stato del mondo.

Questo è il problema del controllo ottimale dell'attenzione: dove guardare, cosa toccare, dove muoversi per ridurre l'incertezza sulle proprietà rilevanti. La struttura bayesiana del framework rende chiaro

perché le azioni ottimali per la riduzione dell'incertezza dipendono dal prior: un agente con prior diverso ha incertezza su dimensioni diverse dello stato del mondo e quindi troverà ottimali azioni diverse.

VIII. Il problema del frame nella memoria: il forgetting ecologico come soluzione

VIII.1 La curva dell'oblio di Ebbinghaus e la struttura matematica del dimenticare

Hermann Ebbinghaus (1885) misurò sistematicamente la velocità con cui le sillabe senza senso apprese vengono dimenticate, scoprendo che l'oblio segue una curva approssimativamente logaritmica:

$$m(t) = a / (\log t + a)$$

dove $m(t)$ è la quota di materiale ricordato al tempo t e a è una costante. La forma più nota è la funzione esponenziale decimata di Wixted e Carpenter (2007):

$$m(t) = e^{-bt^\alpha}$$

con $b > 0$ e $0 < \alpha < 1$. Questa forma di potenza (power function forgetting curve) si adatta ai dati sperimentali meglio della forma esponenziale semplice ($\alpha = 1$) proposta originariamente da Ebbinghaus.

L'interpretazione del dimenticare come soluzione al problema del frame è contro-intuitiva ma formalmente precisa. Anderson e Schooler (1991) proposero che la probabilità di recupero di un'informazione dalla memoria sia ottimizzata per corrispondere alla probabilità che quell'informazione sia rilevante nell'ambiente corrente. Se la rilevanza di un'informazione nel presente dipende da quanto recentemente è stata incontrata (gli eventi recenti sono più rilevanti di quelli remoti), la curva ottimale di ritenzione ha la forma di una funzione di potenza — che corrisponde esattamente alla curva dell'oblio osservata.

Formalmente, se la probabilità che un'informazione sia rilevante nell'ambiente diminuisce come $t^{-\alpha}$ con il tempo trascorso dall'ultimo incontro, allora un sistema ottimizzato per massimizzare la probabilità di recupero della informazione corretta "quando serve" deve retrocedere la probabilità di recupero della stessa quantità — producendo esattamente la funzione di potenza osservata.

La formalizzazione di Anderson e Schooler è:

$$P(\text{rilevante al tempo } t) \propto \sum_{i: \text{ultimo incontro}} (t - t_i)^{-\alpha}$$

dove la somma è su tutte le occorrenze passate dell'informazione. La probabilità di recupero dalla memoria dovrebbe seguire

$$P(\text{recupero}) \propto P(\text{rilevante al tempo } t)$$

producendo la funzione di potenza osservata. Il dimenticare non è un fallimento del sistema di memoria: è la soluzione ottimale al problema del frame per la memoria, nel senso che riduce la probabilità di recupero delle informazioni irrilevanti in misura proporzionale alla loro decrescente probabilità di rilevanza.

IX. Conclusioni: il frame come confine tra computabile e ecologico

Il problema del frame è una delle impossibilità più sottili che i sistemi cognitivi devono affrontare. Non è impossibile in senso logico assoluto — è possibile costruire sistemi che lo risolvono per dominio specifici — ma è impossibile in senso computazionale generale: non esiste un algoritmo di complessità polinomiale che determini le proprietà rilevanti dell'ambiente per un agente arbitrario in un ambiente arbitrario.

I sistemi cognitivi biologici hanno risolto questo problema non computando la risposta ma codificando i prior ecologici nell'architettura stessa dei sistemi percettivi e cognitivi, attraverso la selezione evolutiva. Questi prior — implementati come invarianti percettive (Gibson), come distribuzioni a priori bayesiane, come euristiche veloci e frugali (Gigerenzen), come pattern di oblio ottimizzati (Anderson) — sono la soluzione biologica al problema computazionale impossibile.

Ma questa soluzione è intrinsecamente locale: funziona nell'ambiente evolutivo di adattamento e produce fallimenti sistematici nell'ambiente moderno, che differisce in molte dimensioni critiche dall'EEA. I bias cognitivi, le illusioni percettive, le difficoltà degli esseri umani con il ragionamento statistico e probabilistico: tutti questi fenomeni hanno la stessa struttura formale — prior calibrati su un ambiente che non esiste più applicati a un ambiente nuovo — e sono prevedibili con precisione dai principi dell'ecologia cognitiva.

Questo non è un difetto evolutivo correggibile con più intelligenza o più volontà: è la conseguenza necessaria di aver risolto un problema computazionalmente impossibile attraverso una soluzione localmente ottimale. Comprendere questa struttura è il prerequisito per

comprendere sia i limiti della cognizione umana sia le possibilità di sistemi artificiali che, a differenza di quelli biologici, non sono vincolati dall'eredità evolutiva — ma devono affrontare, anch'essi, la struttura fondamentale del problema del frame.

Monografia IV

Il divario esplicativo in neuroscienza

I tre livelli di Marr, l'explanatory gap di Chalmers, la coscienza come confine e le impossibilità strutturali della riduzione neurale

I. Il paradosso della materia pensante

Nel 1641, René Descartes identificò il problema che ancora oggi — a quasi quattro secoli di distanza e nonostante progressi neuroscientifici straordinari — rimane il più refrattario della filosofia della mente: come può una sostanza materiale, soggetta alle leggi meccaniche della fisica, essere il substrato di un'esperienza soggettiva? Come possono impulsi elettrici che si propagano lungo membrane lipidiche generare il rosso del tramonto, il dolore del lutto, il sapore del vino?

Descartes risolse il paradosso attraverso il dualismo: mente e corpo sono sostanze di natura diversa, la prima immateriale (*res cogitans*), il secondo materiale (*res extensa*). Questa soluzione, però, genera immediatamente un nuovo problema: come interagiscono le due sostanze? Il dualismo cartesiano è la risposta sbagliata a una domanda giusta — e la domanda giusta rimane aperta.

Il progresso della neuroscienza nel Novecento ha consolidato il materialismo come posizione predefinita: la mente è identica a certi processi neurali, e la comprensione della mente richiede la comprensione di questi processi. Ma il materialismo ingenuo — la tesi che una descrizione sufficientemente dettagliata dei processi neurali sia equivalente a una spiegazione della mente — incontra quella che David Chalmers (1995, 1996) ha chiamato "il problema difficile della coscienza" (*the hard problem of consciousness*): anche una spiegazione funzionale completa di tutti i processi cognitivi lascia aperta la questione del perché quegli stessi processi siano accompagnati da esperienza soggettiva.

Questa monografia affronta il divario esplicativo in neuroscienza da due direzioni complementari. La prima è quella della struttura dei livelli di spiegazione, attraverso la gerarchia di David Marr (1982), che mostra come i livelli computazionale, algoritmico e implementazionale siano irriducibili l'uno all'altro pur restando connessi. La seconda è quella della struttura dell'esperienza soggettiva, attraverso l'argomento di Chalmers, che mostra come nessuna descrizione funzionale — per quanto completa — esaurisca il contenuto fenomenico dell'esperienza. Il risultato è una mappa precisa di ciò che la neuroscienza può spiegare e di ciò che, per ragioni strutturali, non potrà mai spiegare nei propri termini.

II. I tre livelli di Marr: la gerarchia dell'analisi computazionale

II.1 La struttura della gerarchia

David Marr, nel libro postumo *Vision* (1982) — terminato durante la malattia e pubblicato dopo la sua morte a trentacinque anni — propose una gerarchia di livelli di analisi per i sistemi di elaborazione dell'informazione che è diventata il framework metodologico dominante nella neuroscienza computazionale. Marr distingue tre livelli distinti, ciascuno dei quali risponde a domande diverse e richiede strumenti concettuali diversi:

Il **livello computazionale** risponde alla domanda: *Cosa calcola il sistema, e perché?* È la specifica dello scopo del sistema — la funzione che implementa e il problema che risolve — indipendentemente dal modo in cui la risolve. Per il sistema visivo, il livello computazionale specifica cosa deve essere calcolato: la struttura tridimensionale della scena, l'identità degli oggetti, la direzione e la velocità del moto. La risposta alla domanda "perché" è in termini ecologici: il sistema visivo deve calcolare queste quantità perché sono biologicamente rilevanti per la navigazione, la predazione, la comunicazione sociale.

Il **livello algoritmico** risponde alla domanda: *Come calcola il sistema, in quali passi, con quale rappresentazione?* È la specifica del metodo di calcolo — l'algoritmo e la struttura dati. Per il sistema visivo, il livello algoritmico specifica come viene calcolata la struttura tridimensionale: attraverso quale sequenza di trasformazioni sui pattern di luminanza retinica, con quali strutture di rappresentazione intermedie (linee, bordi, superfici, volumi), con quale ordine di elaborazione. Algoritmi diversi

possono risolvere lo stesso problema computazionale (calcolare la stessa funzione) — il livello algoritmico specifica la scelta tra questi algoritmi.

Il **livello implementazionale** risponde alla domanda: *Come è fisicamente realizzato il sistema?* È la specifica del substrato materiale — i neuroni, le sinapsi, le correnti ioniche, le molecole di neurotrasmettitore. Per il sistema visivo, il livello implementazionale è la neuroscienza classica: la morfologia delle cellule gangliari della retina, le connessioni tonotopiche del corpo genicolato laterale, la struttura laminare delle aree corticali visive.

II.2 La separabilità dei livelli e le sue conseguenze

Il principio fondamentale della gerarchia di Marr è che i tre livelli sono relativamente separabili: la comprensione di un livello non richiede la comprensione degli altri, e le domande proprie di un livello non sono risolvibili con gli strumenti degli altri livelli.

La separabilità verso il basso è la più intuitiva: sistemi fisicamente diversi (cervelli biologici, reti di silicio, congegni idraulici) possono implementare lo stesso algoritmo, e sistemi con lo stesso substrato fisico (due neuroni con morfologia identica) possono implementare algoritmi diversi. La comprensione del livello implementazionale non determina la comprensione del livello algoritmico: sapere che i neuroni usano potenziali d'azione per trasmettere segnali non dice nulla su quale algoritmo stiano implementando.

La separabilità verso l'alto è meno ovvia ma ugualmente importante: lo stesso algoritmo può essere implementato su substrati fisici diversi (istituzionalizzazione del principio di indifferenza del substrato), e la stessa scopo computazionale può essere raggiunto da algoritmi diversi (pluralismo algoritmico). Questo significa che la domanda "cosa sta calcolando questo sistema?" non è determinata né dalla conoscenza del substrato fisico né dalla conoscenza dell'algoritmo — richiede un'analisi a livello computazionale che è concettualmente autonoma.

Le conseguenze metodologiche sono profonde. La neuroscienza classica lavora prevalentemente al livello implementazionale: registra l'attività di neuroni e reti neurali, studia la connettività anatomica, misura le correnti ioniche. Queste osservazioni sono necessarie ma non sufficienti per la comprensione della mente: anche una descrizione completa di ogni neurone e ogni sinapsi non direbbe nulla sull'algoritmo che quei neuroni implementano né sullo scopo di quell'algoritmo. La psicologia cognitiva lavora prevalentemente al livello algoritmico: studia i processi di elaborazione dell'informazione attraverso tempi di reazione, errori sistematici, effetti di interferenza. Queste osservazioni rispecchiano l'algoritmo ma non il suo scopo né la sua implementazione

fisica. Il livello computazionale — la domanda "perché questo sistema fa quello che fa, quale funzione adempie nella vita dell'organismo" — è il dominio dell'ecologia cognitiva e dell'analisi evolutiva.

II.3 Le irriduzioni formali tra i livelli

La gerarchia di Marr non è soltanto una classificazione descrittiva dei tipi di domande: è una struttura con proprietà formali di irriduzione che si possono enunciare con precisione.

La prima irriduzione è quella del livello computazionale rispetto al livello implementazionale. Formalmente: data una descrizione completa dell'implementazione fisica (neuroni, sinapsi, correnti), non esiste un algoritmo che derivi meccanicamente la specifica computazionale (cosa calcola, perché). Questo segue dal teorema di Rice applicato ai sistemi neurali: il comportamento funzionale di un sistema neurale (la funzione che calcola) è una proprietà semantica del sistema, e le proprietà semantiche non banali dei sistemi di calcolo non sono decidibili dall'esame delle strutture fisiche (che corrispondono alla "macchina di Turing" del sistema).

La dimostrazione è strutturalmente identica a quella del teorema di Rice per i programmi. Sia il sistema neurale S caratterizzato dal suo grafo di connettività $G(S) = (N, E, W)$ dove N è l'insieme dei neuroni, E le connessioni sinaptiche, e W i pesi sinaptici. Sia φ una proprietà semantica non banale del sistema (ad esempio "questo sistema riconosce i volti"). Per il principio di indifferenza del substrato, esistono sistemi con grafo di connettività $G' \neq G(S)$ che calcolano la stessa funzione, e sistemi con grafo $G'' \approx G(S)$ che calcolano funzioni diverse. Non esiste quindi alcuna relazione algoritmica tra $G(S)$ e φ che sia corretta per tutti i sistemi neurali.

La seconda irriduzione è quella del livello algoritmico rispetto al livello implementazionale nella direzione opposta. Data una specifica algoritmica (algoritmo A , strutture dati D), non esiste un mapping canonico al livello implementazionale: l'algoritmo può essere implementato su architetture neurali radicalmente diverse. Il principio di realizzabilità multipla di Putnam (1967) formalizza questa irriduzione: lo stesso stato mentale (e quindi lo stesso algoritmo) può essere realizzato da stati fisici diversi in organismi diversi. Il dolore negli esseri umani è realizzato da certi pattern di attività neurale; il dolore nei polpi (che hanno un sistema nervoso completamente diverso) sarebbe realizzato da pattern completamente diversi. Se dolore = un particolare pattern di attività neurale umana, i polpi non possono provare dolore — una conseguenza inaccettabile che mostra l'insostenibilità del riduzionismo ingenuo.

III. L'explanatory gap: la struttura formale del problema difficile

III.1 La formulazione di Chalmers

Il "problema difficile della coscienza" di Chalmers (1995, 1996) è distinto dal "problema facile" (easy problems). I problemi facili della coscienza includono: spiegare come il cervello integra l'informazione proveniente da diversi sistemi sensoriali; come il cervello controlla il comportamento in modo flessibile; come il cervello genera e mantiene l'attenzione; come distinguiamo il sonno dalla veglia. Questi problemi sono "facili" non nel senso che la loro soluzione sia semplice o vicina, ma nel senso che è chiaro quale tipo di spiegazione sarebbe soddisfacente — una spiegazione funzionale in termini di meccanismi di elaborazione dell'informazione.

Il problema difficile è quello dell'esperienza soggettiva: perché, quando il cervello elabora informazione visiva sulla lunghezza d'onda della luce riflessa da una mela, vi è un'esperienza del rosso? Perché vi è "qualcosa che è come" vedere il rosso, sentire dolore, ascoltare la musica? Thomas Nagel (1974) aveva chiamato questa dimensione dell'esperienza il "carattere qualitativo" o qualia (termine introdotto da Lewis nel 1929 e reso popolare da Dennett). Il problema difficile è quello di spiegare l'esistenza e la natura di questi qualia.

La struttura formale del problema difficile può essere enunciata come un problema di gap esplicativo. Sia $F = \{f_1, f_2, \dots, f\}$ l'insieme completo dei fatti fisici sul sistema nervoso — neurofisiologia, anatomia, dinamica, chimica. Sia E l'insieme dei fatti sull'esperienza soggettiva — i qualia, il carattere fenomenico dell'esperienza. Il gap esplicativo è il fatto che

$E \not\subset \text{closure}(F)$

ossia i fatti sull'esperienza soggettiva non sono derivabili dalla chiusura deduttiva dei fatti fisici. Per qualsiasi fatto fisico $f \in F$ e qualsiasi regola di inferenza, la proposizione "vi è un'esperienza del rosso" non è una conseguenza logica di f (o di qualsiasi congiunzione finita di elementi di F). Questa è la tesi della non-riduzione logica.

III.2 L'argomento dello zombie

Chalmers articola il problema difficile attraverso l'argomento dello zombie filosofico. Un zombie filosofico è un essere fisicamente identico a un essere umano — con lo stesso identico cervello, con le stesse identiche connessioni neurali, con gli stessi identici processi di elaborazione dell'informazione — ma privo di esperienza soggettiva: c'è qualcuno in casa ma non vi è nessuna luce accesa, per usare la metafora di Chalmers.

L'argomento procede in tre passi:

- (1) Gli zombie sono concettualmente possibili: non è contraddittorio immaginare un essere fisicamente identico a un essere umano ma privo di esperienza soggettiva.
- (2) La possibilità concettuale implica la possibilità metafisica: se gli zombie sono concettualmente possibili, non vi è una connessione necessaria tra le proprietà fisiche e le proprietà fenomeniche.
- (3) Se le proprietà fenomeniche non supervenienti necessariamente sulle proprietà fisiche, il fisicalismo è falso: vi sono fatti del mondo che non sono fatti fisici.

La risposta fisicalista più sofisticata è la negazione del passo (2): la possibilità concettuale non implica la possibilità metafisica. Per analogia: H_2O e acqua sono concettualmente distinti (si può immaginare un mondo in cui l'acqua non sia H_2O), ma metafisicamente identici (è metafisicamente impossibile un mondo in cui l'acqua non sia H_2O). Forse la coscienza sta all'attività neurale come l'acqua sta a H_2O : concettualmente distinta, metafisicamente identica.

Chalmers risponde con la nozione di conceivabilità positiva vs. negativa. Per l'acqua e H_2O , l'identità metafisica è "scoperta" empiricamente — è una scoperta della chimica che l'acqua è H_2O . Per la coscienza e l'attività neurale, non è chiaro quale scoperta empirica potrebbe stabilire l'identità: nessuna misura neuroscientifica, per quanto dettagliata, sembra logicamente implicare la presenza di esperienza soggettiva.

La struttura formale di questa risposta usa la nozione di verità a priori e a posteriori. Una proposizione P è necessaria a priori se è vera in tutti i mondi possibili e la sua verità è conoscibile indipendentemente dall'esperienza; è necessaria a posteriori se è vera in tutti i mondi possibili ma la sua verità è conoscibile soltanto empiricamente. L'identità "acqua = H_2O " è necessaria a posteriori. Chalmers argomenta che nessuna proposizione della forma "il processo neurale X = il quale fenomenico Y " è necessaria a posteriori, perché la necessità dell'identità richiederebbe un collegamento a priori tra i termini — ma nessun termine fisico ha collegamento a priori con termini fenomenici.

III.3 L'argomento della conoscibilità: Mary e Frank

Frank Jackson (1982) propose l'argomento dell'esperienza di Mary per illustrare il divario esplicativo da una prospettiva epistemologica. Mary è una scienziata che ha vissuto tutta la sua vita in una stanza in bianco e nero, dove ha studiato completamente tutti i fatti fisici sulla visione dei colori. Conosce la fisica dell'elettromagnetismo, la neurobiologia dei coni e dei bastoncelli, la fisiologia della corteccia visiva, la psicologia della percezione del colore. Sa tutto quello che c'è da sapere, da un punto di vista fisico e funzionale, sulla visione dei colori.

Poi Mary esce dalla stanza e vede il rosso per la prima volta. L'argomento di Jackson è che Mary impara qualcosa di nuovo — impara come è il rosso. Se impara qualcosa di nuovo, allora prima della sua uscita non sapeva tutto — ma per ipotesi sapeva tutti i fatti fisici. Quindi vi sono fatti non fisici — i qualia — che non sono catturati dalla descrizione fisica completa.

L'argomento può essere formalizzato nel linguaggio della teoria dell'informazione. Sia I_{fisico} l'insieme di tutte le proposizioni vere sulla fisica della visione dei colori — la quantità totale di informazione fisica sul sistema. Sia $I_{\text{fenomenico}}$ l'insieme di tutte le proposizioni vere sull'esperienza del rosso. L'argomento di Jackson afferma:

$I_{\text{fenomenico}} \not\subset I_{\text{fisico}}$

ossia l'informazione fenomenica non è contenuta nell'informazione fisica. Formalmente, la quantità di informazione fenomenica che non è recuperabile dall'informazione fisica è

$$I_{\text{fenomenico}} \mid I_{\text{fisico}} = H(I_{\text{fenomenico}}) - I(I_{\text{fenomenico}}; I_{\text{fisico}}) > 0$$

dove H è l'entropia di Shannon e $I(\cdot;\cdot)$ è l'informazione mutua. Il divario esplicativo è questa informazione residua — l'informazione sull'esperienza soggettiva che non è catturata da nessuna descrizione fisica.

IV. Le teorie neuroscientifiche della coscienza e i loro limiti strutturali

IV.1 La Global Workspace Theory di Baars e Dehaene

La Global Workspace Theory (GWT), introdotta da Bernard Baars (1988) e sviluppata nella versione neuroscientifica da Stanislas Dehaene, Jean-Pierre Changeux e collaboratori (1998, 2011), propone che la coscienza corrisponda alla trasmissione globale di informazione attraverso una rete di neuroni distribuiti in varie aree corticali — il "workspace globale".

L'architettura proposta dalla GWT è la seguente. Il cervello contiene una molteplicità di processori specializzati e modulari (per il linguaggio, la visione, l'attenzione, la memoria) che operano in parallelo e autonomamente. Un sottoinsieme di questi processori è connesso a un "workspace globale" — una rete di neuroni prefrontali e parietali con connessioni a lungo raggio. Quando un'informazione accede al workspace globale — grazie a una competizione tra processori per l'attenzione — essa diventa disponibile globalmente a tutti gli altri processori, e questo accesso globale costituisce l'esperienza cosciente.

La formalizzazione neuroscientifica di Dehaene usa il concetto di "ignition": l'attivazione improvvisa e sostenuta di un'ampia rete fronto-parietale, caratterizzata da potenziali event-related (ERP) tardivi (P300, N400) e da una cascata di attivazioni che si diffonde dalla corteccia sensoriale alle aree frontali e parietali. Neuroscientificamente, la GWT è una delle teorie meglio supportate empiricamente: l'ignition fronto-parietale è osservata in risposta a stimoli coscienti e assente per stimoli subliminali, controllando per l'intensità dello stimolo.

Tuttavia, la GWT incontra un limite strutturale preciso rispetto al problema difficile. La GWT spiega quale informazione è cosciente (quella che accede al workspace globale) e quando (quando scatta l'ignition), ma non spiega perché l'accesso globale sia accompagnato da esperienza soggettiva. Questo è esattamente un problema facile, nel senso di Chalmers: spiega il correlato funzionale della coscienza — il meccanismo che distingue l'elaborazione cosciente da quella inconscia — senza spiegare il carattere fenomenico dell'esperienza.

La struttura formale del limite è la seguente. La GWT stabilisce una relazione di covarianza:

$$P(\text{ignition} \mid \text{stimolo cosciente}) > P(\text{ignition} \mid \text{stimolo non cosciente})$$

Ma questa covarianza è compatibile con l'ipotesi zombie: uno zombie filosofico mostrerebbe esattamente la stessa covarianza — lo stesso pattern di ignition fronto-parietale — in assenza di qualsiasi esperienza soggettiva. La GWT è una teoria dei correlati neurali della coscienza (NCC), non una teoria della coscienza stessa.

IV.2 La Teoria dell'Informazione Integrata di Tononi

La Teoria dell'Informazione Integrata (IIT), proposta da Giulio Tononi (2004, 2008) e sviluppata collaborativamente con Christof Koch, propone che la coscienza sia identica a un certo tipo e quantità di informazione integrata nel sistema — non soltanto la quantità di informazione, ma la quantità di informazione che non può essere scomposta nelle informazioni delle parti.

La misura centrale della IIT è il parametro Φ (phi), che quantifica l'informazione integrata di un sistema rispetto alla sua miglior decomposizione in parti indipendenti. La definizione formale di Φ in un sistema dinamico discreto (Tononi, 2008) è:

$$\Phi = I(X_t; X_{t-1}) - \max_{\text{bipartizioni } M} I(X_t^M; X_{t-1}^M)$$

dove $I(X_t; X_{t-1})$ è l'informazione mutua tra lo stato attuale del sistema e il suo stato passato, e $\max_{\text{bipartizioni } M} I(X_t^M; X_{t-1}^M)$ è la massima informazione mutua ottenibile da qualsiasi bipartizione del sistema in due sottosistemi indipendenti. Φ misura quanto il sistema "fa di più" rispetto alla somma delle sue parti in termini di capacità di trasmettere informazione attraverso il tempo.

La versione più recente (IIT 3.0, Oizumi, Albantakis e Tononi, 2014) usa la "distanza di Earth Mover" (distanza di Wasserstein tra distribuzioni di probabilità) per misurare la quantità di informazione integrata:

$$\Phi = \min_{\text{bipartizioni } M} W_1(p(X_t | X_{t-1}), p^M(X_t | X_{t-1}))$$

dove W_1 è la distanza di Wasserstein di ordine 1 (la distanza dell'"Earth Mover") e p^M è la distribuzione indipendente per la bipartizione M .

La IIT ha proprietà formali notevoli. Primo, è genuinamente quantitativa: sistemi diversi hanno valori diversi di Φ , e il livello di coscienza è proporzionale a Φ . Secondo, è intrinseca: il valore di Φ dipende soltanto dalla struttura causale interna del sistema, non dalla sua funzione o dal suo comportamento osservabile. Terzo, è realizzabilità-multiplo: sistemi fisicamente diversi possono avere lo stesso Φ e quindi lo stesso livello di coscienza. Quarto, ha conseguenze controverse: una rete di fotorivelatori in griglia può avere $\Phi > 0$, il che implica che potrebbe avere un qualche grado di coscienza — una predizione che molti trovano difficile da accettare.

Il limite della IIT rispetto al problema difficile è diverso da quello della GWT: la IIT non propone soltanto un NCC ma propone un'identità — la coscienza è l'informazione integrata. Ma questa mossa — sostituire la correlazione con l'identità — non risolve il problema difficile, la sposta. La domanda "perché l'accesso globale è accompagnato da esperienza?"

diventa "perché l'informazione integrata è accompagnata da esperienza?". La risposta della IIT — "per definizione, la coscienza è informazione integrata" — è un'assunzione, non una spiegazione. Il problema difficile persiste: perché dovrebbe esserci un'esperienza soggettiva associata a qualsiasi valore di Φ ?

IV.3 La Teoria dell'Ordine Superiore e il problema della regressione

Le teorie dell'ordine superiore (Higher Order Theories, HOT) propongono che uno stato mentale sia cosciente quando è oggetto di un altro stato mentale di ordine superiore — una rappresentazione del primo stato. David Rosenthal (1997, 2005) ha sviluppato la versione più sistematica: uno stato mentale M è cosciente se e solo se il sistema ha un pensiero di ordine superiore (HOT) che rappresenta M come proprio stato corrente.

La struttura formale è la seguente. Sia $S(t)$ lo stato del sistema al tempo t e $HOT(M)$ la rappresentazione di M come stato corrente del sistema. La coscienza è la relazione:

$$C(M) \Leftrightarrow HOT(M)$$

La coscienza del colore rosso richiede che vi sia uno stato M che rappresenta il rosso (un NCC del rosso al livello primario) e un HOT che rappresenta M come stato attuale del sistema (una meta-rappresentazione dello stato primario).

Le teorie HOT affrontano un problema di regressione potenziale: se un stato è cosciente perché è oggetto di un HOT, il HOT stesso è cosciente? Se sì, richiede un HHOT (Higher Higher Order Thought) — e così via per regressione infinita. Rosenthal risponde che il HOT che rende cosciente lo stato primario non deve essere esso stesso cosciente — può essere uno stato mentale inconscio di ordine superiore. Ma questa risposta solleva una domanda diversa: come un pensiero inconscio può rendere cosciente uno stato che altrimenti non lo sarebbe?

Il limite strutturale delle teorie HOT rispetto al problema difficile è il medesimo delle altre teorie funzionali: spiegano il meccanismo che determina quali stati sono coscienti, ma non spiegano perché la presenza di quel meccanismo sia accompagnata da esperienza soggettiva. Anche qui, lo zombie filosofico con esattamente la stessa architettura HOT sarebbe funzionalmente indistinguibile da un essere cosciente.

V. La struttura formale del divario esplicativo

V.1 Il divario come irriducibilità a-priori

Il divario esplicativo può essere formalizzato in modo più preciso attraverso il concetto di implicazione a priori e a posteriori. Un fatto P implica a priori un fatto Q se la verità di Q segue logicamente dalla verità di P, senza richiedere osservazioni empiriche aggiuntive. Un fatto P implica a posteriori Q se la connessione tra P e Q è stabilita empiricamente — è una scoperta, non una derivazione logica.

Chalmers (2010) formalizza questo come segue. Sia il sistema neurale fisicamente descritto da T (una teoria fisica completa del sistema) e l'esperienza fenomenica descritta da Q. La relazione di implicazione a priori è:

$$T \rightarrow_{\text{a priori}} Q$$

se la verità di T, congiuntamente alle definizioni dei termini teorici, implica logicamente Q senza osservazioni empiriche aggiuntive.

Per i fatti funzionali — il fatto che il sistema discrimini lunghezze d'onda diverse, il fatto che reagisca selettivamente a certi stimoli visivi, il fatto che produca certe risposte verbali riguardo al colore — l'implicazione a priori vale:

$$T \rightarrow_{\text{a priori}} \text{"il sistema discrimina 700nm da 500nm"}$$

Per i fatti fenomenici — il fatto che vi sia un'esperienza del rosso — l'implicazione a priori non vale:

$$T \not\rightarrow_{\text{a priori}} \text{"vi è un'esperienza del rosso"}$$

Questo è il divario esplicativo nella sua formulazione più precisa: i fatti fisici implicano a priori i fatti funzionali ma non i fatti fenomenici. Vi è sempre una lacuna tra la descrizione fisica completa e la descrizione fenomenica — non per ignoranza, ma per la struttura stessa delle proposizioni fisiche e fenomeniche.

V.2 La zombie completeness e i suoi limiti

Jackson (1998) ha proposto un test formale per valutare la completezza di una teoria della mente: la "zombie completeness". Una teoria T della coscienza è zombie-completabile se, dato il modello fisico completo del cervello M(T) previsto da T e le proprie assunzioni, è possibile derivare

l'assenza di esperienza soggettiva come implicazione logica — ossia se il mondo fisico previsto da T è compatibile con l'assenza di qualia.

Il test di zombie-completability mostra che le teorie funzionaliste della coscienza (GWT, IIT nel caso la sua identità non sia accettata, HOT) sono zombie-completibili: il mondo fisico che queste teorie descrivono è compatibile con l'ipotesi zombie (un sistema con la stessa struttura funzionale ma senza esperienza). Questo non è un difetto contingente delle teorie — è la conseguenza strutturale del fatto che esse descrivono soltanto la struttura funzionale della mente, non il suo carattere fenomenico.

L'unica teoria non zombie-completibile sarebbe una teoria che identifica la coscienza non con una struttura funzionale ma con una proprietà intrinseca della materia — una "proto-fenomenicità" che è presente in ogni struttura fisica e si integra in esperienza soggettiva quando la struttura ha le proprietà appropriate. Questa è la posizione del pansichismo, nella versione del costitutivismo fenomenico di Chalmers o nella versione della proto-coscienza di Goff (2017). Il pansichismo risolve il problema difficile trasferendo la fenomenicità al livello fisico fondamentale, ma a costo di postulare proprietà intrinseche della materia che sono per definizione non osservabili e non riducibili ai fatti strutturali della fisica.

VI. La neuroscienza della coscienza: i correlati neurali (NCC) e il loro status

VI.1 La metodologia degli NCC

Crick e Koch (1990) proposero il programma empirico degli NCC (Neural Correlates of Consciousness): identificare i minimi meccanismi neurali sufficienti per una data esperienza cosciente. Questo programma è distinto dalle teorie della coscienza — non propone una spiegazione di perché i meccanismi neurali siano accompagnati da esperienza, ma si limita a identificare quali meccanismi sono necessari e sufficienti per produrre dati fenomenologici in soggetti umani.

La metodologia degli NCC usa il contrasto tra condizioni sperimentali con e senza esperienza cosciente del medesimo stimolo. Le tecniche principali sono:

Binocular rivalry: quando un occhio riceve un'immagine e l'altro un'immagine diversa, la percezione alterna tra le due immagini ogni pochi secondi. Lo stesso stimolo fisico produce esperienze alternanti,

permettendo di confrontare l'attività neurale durante la percezione di ciascuna immagine a parità di input sensoriale.

Change blindness e inattentional blindness: cambiamenti dell'immagine o oggetti inattesi non vengono percepiti se l'attenzione è diretta altrove. La differenza nell'attività neurale tra le condizioni "percepisce il cambiamento" e "non percepisce il cambiamento" isolano i correlati dell'attenzione e della coscienza.

Masking backward: uno stimolo breve può essere reso non cosciente da uno stimolo successivo (la maschera) che sovrascrive la sua rappresentazione. Confrontando le condizioni con e senza masking a parità di stimolo si isolano i correlati dell'accessibilità cosciente.

I risultati convergono su un insieme di aree e meccanismi associati alla coscienza: la corteccia prefrontale dorsolaterale e la corteccia parietale posteriore (la rete fronto-parietale della GWT), i meccanismi di sincronizzazione nella banda gamma (30–80 Hz), le connessioni a lungo raggio tra aree corticali. Questi sono i NCC nel senso di Crick e Koch — ma rimane aperto se siano sufficienti o soltanto necessari, e rimane completamente aperto il problema difficile.

VI.2 I marker quantitativi della coscienza: il PCI di Casali et al.

Un risultato notevole nella ricerca empirica sulla coscienza è l'indice di complessità perturbazionale (Perturbational Complexity Index, PCI) sviluppato da Casali et al. (2013) come marker quantitativo del livello di coscienza.

Il PCI combina due misure: la complessità della risposta corticale a una perturbazione (stimolazione magnetica transcranica, TMS) e la sua distribuzione spaziale. Formalmente, il PCI è definito come

$$\text{PCI} = K_{\text{LZ}}(A_{\{\text{TMS}\}}) / N_{\text{bits}}$$

dove K_{LZ} è la complessità di Lempel-Ziv della sequenza binaria A_{TMS} ottenuta dalla compressione delle onde EEG registrate dopo la TMS (la risposta elettrica è binarizzata con soglia alla mediana e la sequenza binaria risultante è compressa con l'algoritmo di Lempel-Ziv), e N_{bits} è il numero di bit nella sequenza originale (per normalizzare rispetto alla lunghezza della sequenza).

Il PCI ha diverse proprietà notevoli:

In soggetti svegli in stato di veglia, $\text{PCI} \approx 0.44\text{--}0.70$.

In soggetti addormentati in fase NREM (sonno profondo non-REM, privo di sogni), $PCI \approx 0.20-0.35$.

In soggetti in anestesia con ketamina o propofol, $PCI \approx 0.15-0.30$.

In pazienti vegetativi (stati di minima coscienza confermati), $PCI \approx 0.20-0.35$.

Il PCI distingue con elevata accuratezza ($>90\%$) i soggetti coscienti dai soggetti non coscienti, anche in pazienti per i quali la valutazione clinica è difficoltosa (stati vegetativi, locked-in syndrome).

La struttura matematica del PCI come misura della complessità algoritmica della risposta corticale è notevole: la complessità di Lempel-Ziv è una stima computabile della complessità di Kolmogorov, che abbiamo visto nella Monografia XVI del ciclo precedente come misura fondamentale dell'informazione strutturale. Il PCI cattura quindi la quantità di struttura computazionale nella risposta del cervello a una perturbazione esterna — una misura genuinamente informativa del livello di integrazione corticale.

Tuttavia, il PCI è un NCC, non una spiegazione della coscienza: misura un correlato — la complessità della risposta corticale — senza spiegare perché quella complessità sia accompagnata da esperienza soggettiva.

VII. Il problema della localizzazione e la distribuzione dell'esperienza

VII.1 Il problema dell'unità dell'esperienza

Uno degli aspetti più sorprendenti dell'esperienza cosciente è la sua unità: in ogni momento, vi è un'esperienza unificata che integra informazione proveniente da tutti i sistemi sensoriali, dalla memoria e dall'immaginazione in un unico "campo di coscienza". Questa unità è paradossale rispetto alla struttura distribuita del cervello: l'elaborazione visiva avviene in V1, V2, V4 e IT (aree posteriori); il linguaggio in Broca e Wernicke (aree frontali e temporali sinistre); l'emozione nell'amigdala, nel nucleo accumbens e nella corteccia prefrontale orbitofrontale. Come si integra questa elaborazione distribuita in un'esperienza unitaria?

Il problema dell'unità è una versione del problema del binding analizzato nella Monografia I. La differenza è che qui il binding non riguarda le feature di un singolo oggetto percettivo, bensì l'integrazione di contenuti cognitivi distribuiti in aree completamente diverse del

cervello — il contenuto dell'esperienza visiva con il contenuto dell'esperienza propriocettiva con il contenuto del pensiero verbale con il contenuto del ricordo.

L'evidenza dei pazienti callosotomizzati (con sezionamento del corpo calloso, il fascio di fibre che connette i due emisferi) è particolarmente istruttiva. Negli esperimenti di Gazzaniga e Sperry (1960s–70s), i pazienti con cervello diviso mostrano due sistemi cognitivi semi-indipendenti che possono avere credenze diverse, intenzioni diverse e persino risposte emotive diverse allo stesso stimolo. Quale dei due sistemi "ha" l'esperienza cosciente? La risposta più parsimoniosa è che entrambi hanno esperienze, ma si tratta di due esperienze semi-indipendenti — non un'unica esperienza divisa — il che suggerisce che l'unità dell'esperienza dipende dall'integrità delle connessioni tra emisferi.

VII.2 Il cervelletto e il paradosso della non-coscienza

Il cervelletto contiene più della metà di tutti i neuroni del cervello — circa 69 miliardi su un totale di 86 miliardi — organizzati in una struttura altamente regolare e cristallina. Il cervelletto elabora informazione complessa sul movimento, la postura, la coordinazione motoria e (recentemente si è scoperto) alcune funzioni cognitive. Eppure, danni al cervelletto tipicamente non producono perdita di coscienza: i pazienti con lesioni cerebellari estese mantengono un pieno livello di coscienza (sebbene la coordinazione motoria sia gravemente compromessa).

Questo è paradossale rispetto alle teorie della coscienza basate sull'elaborazione dell'informazione: il cervelletto elabora enormi quantità di informazione in modo coordinato e temporalmente preciso, ma questa elaborazione non è accompagnata da coscienza (o è accompagnata da un livello di coscienza trascurabilmente basso). Perché?

Tononi e Koch (2015) propongono la risposta nel contesto della IIT: il cervelletto, nonostante il numero enorme di neuroni, ha un'architettura che produce valori bassi di Φ . La struttura cerebellare è quella di un "feed-forward" puro: le cellule granulari proiettano alle cellule di Purkinje attraverso le fibre parallele, le cellule di Purkinje proiettano ai nuclei cerebellari, senza connessioni a lungo raggio ricorrenti tra le diverse aree cerebellari. Questa struttura produce informazione integrata bassa (le parti del cervelletto elaborano relativamente indipendentemente), e quindi basso Φ , e quindi bassa coscienza. La corteccia cerebrale, al contrario, è ricca di connessioni ricorrenti a lungo raggio che producono alto Φ .

Che la IIT abbia ragione o meno nel dettaglio, il paradosso del cervelletto mostra che la quantità di elaborazione dell'informazione non è il fattore determinante della coscienza: è la struttura di quella elaborazione — in qualche dimensione che le diverse teorie cercano di catturare (informazione globale per la GWT, informazione integrata per la IIT) — che è rilevante.

VIII. La fenomenologia come dato primario e il problema dell'accesso

VIII.1 La fenomenologia come metodologia

Edmund Husserl e la tradizione fenomenologica proposero che l'esperienza soggettiva — la coscienza in prima persona — debba essere studiata direttamente, come dato primario, attraverso la "riduzione fenomenologica" (*epoché*): sospendere i giudizi sull'esistenza esterna del mondo e descrivere l'esperienza così come si dà all'osservatore, nella sua struttura intenzionale e nei suoi contenuti qualitative.

La fenomenologia come metodologia ha prodotto una ricca descrizione della struttura dell'esperienza. Husserl ha analizzato la struttura temporale della coscienza — il "presente vivente" come struttura tridimensionale che integra il ritenuto (appena trascorso), il prototenuto (attuale) e la protensione (imminente). Merleau-Ponty ha analizzato la struttura corporea dell'esperienza — il corpo non come oggetto percepito ma come soggetto percipiente, il "corpo proprio" che è la condizione di possibilità di ogni percezione spaziale. Heidegger ha analizzato la struttura esistenziale dell'essere-nel-mondo — la *Sorge* (cura) come struttura fondamentale del *Dasein*, il modo in cui il mondo si apre attraverso gli strumenti e le relazioni pratiche.

La sfida per la neuroscienza è come integrare questo dato di prima persona — la descrizione fenomenologica dell'esperienza — con i dati di terza persona della neuroscienza sperimentale. Varela, Thompson e Rosch (1991) hanno proposto il programma della "neurofenomenologia": usare le descrizioni fenomenologiche come guide per identificare i correlati neurali dell'esperienza, e usare i correlati neurali come vincoli per affinare le descrizioni fenomenologiche. Questo programma è metodologicamente interessante ma incontra il limite fondamentale del divario esplicativo: anche la correlazione più precisa tra descrizione fenomenologica e attività neurale non spiega il carattere fenomenico dell'esperienza.

VIII.2 Il problema dell'accesso e la distinzione fenomenica-di-accesso

Block (1995) ha proposto una distinzione fondamentale tra due sensi di coscienza: la coscienza di accesso (A-consciousness) e la coscienza fenomenica (P-consciousness).

La coscienza di accesso è la disponibilità funzionale di un'informazione al ragionamento, al controllo del comportamento e al reporto verbale. Uno stato è A-cosciente se la sua rappresentazione è disponibile come premessa per il ragionamento, per le decisioni volontarie e per il controllo verbale del comportamento.

La coscienza fenomenica è il carattere qualitativo dell'esperienza — i qualia. Uno stato è P-cosciente se vi è qualcosa che è come trovarsi in quello stato.

Block argomenta che le due nozioni sono distinte: vi possono essere stati A-coscienti senza P-coscienza (l'elaborazione robotica, lo zombie filosofico) e forse vi possono essere stati P-coscienti senza A-coscienza (alcune esperienze periferiche molto fuggevoli che non vengono mai riportate).

La distinzione di Block ha conseguenze per la neuroscienze della coscienza: la maggior parte delle misure sperimentali — il reporto verbale del soggetto, le risposte comportamentali, i potenziali ERP — misurano la coscienza di accesso, non la coscienza fenomenica. La GWT è esplicitamente una teoria della coscienza di accesso. Il problema difficile riguarda la coscienza fenomenica — la domanda di perché l'accesso globale sia accompagnato da qualia.

IX. Il divario esplicativo come struttura necessaria, non contingente

IX.1 L'argomento dalla necessità del divario

Chalmers (2010) ha formalizzato l'argomento che il divario esplicativo non è contingente — non è il risultato della nostra ignoranza attuale che la neuroscienza del futuro colmerà — ma è strutturalmente necessario, derivante dalla natura stessa dei concetti fisici e fenomenici.

L'argomento usa la nozione di concetti fenomenici. Un concetto fenomenico è un concetto che si riferisce a una qualità dell'esperienza — il concetto del quale del rosso, del quale del dolore, del quale del

gusto del vino. I concetti fenomenici hanno una caratteristica peculiare: sono acquisiti direttamente attraverso l'esperienza (non attraverso la descrizione funzionale), e si riferiscono alle qualità dell'esperienza attraverso un meccanismo di presentazione diretta invece che attraverso una descrizione.

Chalmers argomenta che qualsiasi teoria fisica — per quanto completa — usa soltanto concetti strutturali: concetti che si riferiscono alle relazioni tra le entità, non alle loro proprietà intrinseche. La fisica quantistica descrive le proprietà degli elettroni attraverso relazioni — lo spin, la carica, la massa sono definiti attraverso le relazioni che gli elettroni intrattengono con altri sistemi, non attraverso proprietà intrinseche dell'elettrone in sé. I concetti fenomenici, al contrario, si riferiscono a proprietà intrinseche dell'esperienza — la qualità intrinseca del rosso, indipendentemente da qualsiasi relazione.

Dato che la fisica usa soltanto concetti strutturali e i concetti fenomenici si riferiscono a proprietà intrinseche non strutturali, vi è una lacuna concettuale necessaria tra descrizione fisica e descrizione fenomenica:

Proposizioni strutturali $\neq$ Proposizioni sulle proprietà intrinseche

Il divario esplicativo rispecchia questa lacuna concettuale — non è colmabile in linea di principio da nessuna scoperta neuroscientifica futura, a meno che la neuroscienza non trovi un modo di esprimere le proprietà intrinseche dell'esperienza in termini strutturali — ma allora non sarebbe più una descrizione dell'esperienza stessa.

IX.2 Le conseguenze per il programma neuroscientifico

Le conseguenze dell'impossibilità del divario esplicativo per il programma neuroscientifico sono di tre tipi.

Il primo è metodologico: la ricerca sui NCC è un programma legittimo e importante, ma deve essere accompagnata dalla consapevolezza che i NCC non sono spiegazioni — sono correlati. La scoperta dei NCC descrive il paesaggio empirico della coscienza, ma non risponde alla domanda di perché quei meccanismi neurali siano accompagnati da esperienza. Questa domanda richiede un'analisi concettuale che è al di là del metodo sperimentale standard.

Il secondo è teorico: le teorie della coscienza che propongono identità tra stati neurali e stati coscienti (fisicalismo dell'identità dei tipi, materialismo eliminativo) devono confrontarsi con il fatto che l'identità non è ricavabile a priori dalle descrizioni fisiche — richiede

un'assunzione filosofica aggiuntiva che non può essere confermata empiricamente.

Il terzo è euristico: il divario esplicativo suggerisce che la coscienza potrebbe richiedere un'espansione della nostra ontologia — l'ammissione di proprietà del mondo che non sono catturate dalla fisica strutturale. Il panpsichismo, il dualismo delle proprietà e il pamprotopsichismo di Chalmers sono diverse proposte in questa direzione. Nessuna è priva di difficoltà; tutte affrontano il divario esplicativo in modo più diretto delle teorie puramente funzionaliste.

X. Conclusioni: i tre livelli e il quarto

Il percorso di questa monografia ha mostrato che la gerarchia di Marr, già articolata in tre livelli irriducibili l'uno all'altro, è incompleta rispetto al problema della coscienza: vi è un quarto livello — il livello fenomenico — che non è catturato da nessuno dei tre livelli di Marr e che è strutturalmente irriducibile ad essi.

Il livello computazionale descrive cosa calcola il sistema — quale funzione cognitiva adempie — senza descrivere come è vissuta quella funzione dall'interno. Il livello algoritmico descrive il meccanismo — l'algoritmo e la struttura dati — senza descrivere il carattere dell'esperienza che quel meccanismo produce. Il livello implementazionale descrive il substrato fisico — i neuroni, le sinapsi, le correnti — senza toccare la questione dell'esperienza soggettiva.

Il quarto livello — il livello fenomenico — è la descrizione dell'esperienza soggettiva in prima persona: il quale del rosso, il dolore acuto, la qualità specifica del gusto del caffè. Questo livello non è riducibile ai tre livelli di Marr per ragioni che abbiamo visto essere strutturali, non contingenti. Il divario esplicativo è la distanza tra il terzo livello (implementazionale) e il quarto livello (fenomenico) — una distanza che la neuroscienza, per quanto avanzata, non può colmare rimanendo nel proprio vocabolario descrittivo.

Questo non è un fallimento della neuroscienza: è la specificazione esatta di ciò che la neuroscienza può e non può spiegare. La neuroscienza può spiegare completamente i livelli computazionale, algoritmico e implementazionale della mente — una conquista intellettuale di straordinaria portata. Non può spiegare il livello fenomenico perché quello si trova dall'altra parte di un confine che non è empirico ma concettuale. Come le proposizioni indecidibili di Gödel non sono un fallimento della logica ma la specificazione esatta dei confini della logica

formale, il divario esplicativo non è un fallimento della neuroscienza ma la specificazione esatta dei confini della spiegazione in terza persona.

Le monografie successive del ciclo esploreranno come i sistemi biologici hanno sviluppato, nel corso dell'evoluzione, le strutture che producono questo quarto livello — e perché, nelle analisi dei cicli evolutivi e dello sviluppo, il problema fenomenico rimane il confine più ostinato tra il vivente e il semplicemente fisico.

Monografia V

I limiti della plasticità sviluppativa

Periodi critici, path dependence neurale, la teoria epigenetica e l'irreversibilità strutturale dello sviluppo

I. Il paradosso della finestra chiusa

Tra i fenomeni più sorprendenti della biologia dello sviluppo vi è il fatto che la capacità di acquisire certe competenze — il linguaggio, la percezione del tono musicale assoluto, la visione binoculare — non è uniformemente distribuita nella vita di un organismo, ma è concentrata in finestre temporali specifiche, dette periodi critici o sensibili, al di là delle quali quella capacità non può più essere acquisita con la stessa facilità, e spesso non può essere acquisita affatto. Un bambino esposto a una lingua naturale nei primi anni di vita la acquisirà senza sforzo e con competenza nativa; un adulto che apprende quella stessa lingua per la prima volta non raggiungerà mai lo stesso livello di padronanza, per quanto intenso sia lo studio.

Questo fenomeno è paradossale dal punto di vista di un ottimizzatore ingenuo: se acquisire il linguaggio è adattativo, perché la capacità di acquisirlo diminuisce proprio quando il sistema cognitivo — più potente, più ricco di conoscenze pregresse, più capace di astrazione — dovrebbe avere gli strumenti migliori per farlo? La risposta rivela qualcosa di strutturalmente profondo sull'organizzazione dei sistemi biologici: la plasticità sviluppativa non è un lusso aggiunto al sistema maturo, bensì il meccanismo attraverso cui il sistema maturo viene costruito. Una volta che la struttura è costruita, la plasticità stessa è il componente che viene dismesso — perché la sua conservazione interferirebbe con la stabilità della struttura acquisita.

Questa monografia sviluppa la teoria matematica dei periodi critici, analizzando le strutture formali della path dependence neurale — il fatto che lo sviluppo del sistema nervoso è irreversibilmente determinato dalla sequenza delle esperienze precoci — e le sue conseguenze per la comprensione dei limiti strutturali dell'apprendimento adulto. Il filo conduttore è la dimostrazione che la chiusura dei periodi critici non è un difetto evolutivo ma una necessità strutturale: il prezzo biologico della stabilità delle rappresentazioni costruite durante lo sviluppo.

II. La scoperta di Hubel e Wiesel: i periodi critici della visione

II.1 La privazione monoculare e la plasticità corticale

David Hubel e Torsten Wiesel, nel loro lavoro premiato con il Nobel (1981), scoprirono che l'architettura funzionale della corteccia visiva primaria del gatto — le colonne di dominanza oculare, la selettività per l'orientamento, la preferenza per certi tipi di stimoli — non è completamente predeterminata geneticamente, ma richiede l'esperienza visiva per consolidarsi. La scoperta fondamentale emerge dall'esperimento di privazione monoculare (MD): cucire chiuso un occhio di un gatto neonato durante le prime settimane di vita produce effetti permanenti e drammatici sulla corteccia visiva — non nell'occhio privato (che rimane anatomicamente e fisiologicamente normale) ma nel cervello.

Nell'adulto normale, la corteccia visiva primaria (V1) è organizzata in colonne di dominanza oculare alternanti: bande verticali di neuroni che rispondono preferenzialmente all'occhio sinistro e bande che rispondono preferenzialmente all'occhio destro. Questa organizzazione alternata emerge durante il periodo critico neonatale attraverso la competizione sinaptica tra le proiezioni dei due occhi al nucleo genicolato laterale (LGN) e da questo alla corteccia.

Nella privazione monoculare durante il periodo critico, l'occhio aperto "conquista" più territorio corticale a spese dell'occhio chiuso: le colonne dell'occhio aperto si espandono, quelle dell'occhio chiuso si restringono. La conseguenza funzionale è l'ambliopia — la riduzione permanente dell'acuità visiva nell'occhio privato, nonostante la struttura oculare sia normale. L'ambliopia persiste nell'adulto anche se l'occhio viene riaperto: la finestra è chiusa.

Crucialmente, la privazione monoculare dello stesso animale adulto — dopo la chiusura del periodo critico — non produce alcun effetto sull'organizzazione corticale. Il cervello adulto è impermeabile alla stessa manipolazione che il cervello giovane è esquisitamente sensibile.

II.2 La struttura matematica del periodo critico: il modello di competizione hebbiana

La formalizzazione matematica dei periodi critici richiede un modello della competizione sinaptica durante lo sviluppo. Il framework più generale è quello delle regole di apprendimento hebbiane con competizione — la versione biologica del problema di ottimizzazione con vincolo che emerge dalla teoria dell'apprendimento.

Sia $W_{ij}(t)$ il peso sinaptico dalla proiezione j (ad esempio l'assone dal nucleo genicolato che rappresenta il pixel j del campo visivo) al neurone corticale i , al tempo t . La regola di Hebb non normalizzata è:

$$dW_{ij}/dt = \eta \cdot r_i(t) \cdot x_j(t)$$

dove $r_i(t)$ è il tasso di scarica del neurone postsinaptico i , $x_j(t)$ è l'attività dell'ingresso presinaptico j , e $\eta > 0$ è il tasso di apprendimento. Questa regola, nella sua forma pura, porta all'instabilità: i pesi crescono senza limite, amplificando le correlazioni positive in modo esplosivo.

La regola di Oja (1982) stabilizza la regola di Hebb introducendo una normalizzazione che mantiene la norma del vettore dei pesi costante:

$$dW_{ij}/dt = \eta \cdot r_i(t) \cdot [x_j(t) - r_i(t) W_{ij}(t)]$$

Il termine di normalizzazione $-\eta \cdot r_i(t)^2 \cdot W_{ij}(t)$ introduce una competizione: quando il neurone postsinaptico i scarica fortemente, i pesi verso tutti gli ingressi vengono ridotti proporzionalmente all'entità dell'attività sinaptica $W_{ij}(t)$. Questo meccanismo produce la competizione "winner-take-all" che organizza le colonne di dominanza oculare.

Per il caso dei due occhi che competono per il territorio corticale, si possono definire W_L e W_R come i pesi totali dai due occhi (sinistro e destro) a un neurone corticale. La dinamica della competizione è:

$$\begin{aligned} dW_L/dt &= \eta [C_{LL} W_L + C_{LR} W_R - \lambda(W_L + W_R) W_L] \\ dW_R/dt &= \eta [C_{RR} W_R + C_{RL} W_L - \lambda(W_L + W_R) W_R] \end{aligned}$$

dove C_{LL} e C_{RR} sono le autocorrelazioni degli ingressi dell'occhio sinistro e destro rispettivamente, C_{LR} e C_{RL} sono le correlazioni

incrociate tra i due occhi, e λ è il parametro di normalizzazione competitiva.

Nel caso binoculare normale ($C_{LL} = C_{RR} = C_1$, $C_{LR} = C_{RL} = C_2$ con $C_1 > C_2$), il punto fisso stabile della dinamica è uno stato binoculare:

$$W_{L^*} = W_{R^*} = C_1 / (\lambda(C_1 + C_2))$$

Nella privazione monoculare ($C_{LL} = C_1$, $C_{LR} = C_{RL} = C_2 \approx 0$), il punto fisso stabile è uno stato monoculare:

$$W_{L^*} = C_1/\lambda, W_{R^*} = 0$$

L'occhio privato perde tutte le sue connessioni corticali. La velocità di convergenza verso questo punto fisso è determinata dall'autovalore dominante del sistema linearizzato attorno al punto fisso:

$$\lambda_{\max} = \eta(C_1 - C_2)$$

Quando $C_1 \gg C_2$ (gli ingressi dei due occhi sono scorrelati), la convergenza è rapida e il sistema si orienta monoculare velocemente. Quando $C_1 \approx C_2$ (gli ingressi sono altamente correlati), la convergenza è lenta e il sistema rimane più facilmente binoculare.

II.3 Il ruolo del fattore BDNF e la biologia molecolare dei periodi critici

La domanda biologica fondamentale è: cosa apre e chiude i periodi critici? Decenni di ricerca hanno identificato diversi meccanismi molecolari, ma la struttura più chiara emerge dal ruolo dei fattori neurotrofici — in particolare il BDNF (Brain-Derived Neurotrophic Factor) — e dell'equilibrio eccitatorio-inibitorio nella corteccia.

Il modello attuale, sviluppato principalmente da Takao Hensch e collaboratori, propone che i periodi critici siano regolati dall'attività delle cellule di Parvalbumin — interneuroni inibitori veloci che formano sinapsi somatiche (sinapsi sul corpo cellulare) sui neuroni piramidali. L'organizzazione dei neuroni di Parvalbumin in "reti perisomatic" — strutture extracellulari dette reti pericellulari (perineuronal nets, PNN) — stabilizza le sinapsi che i neuroni piramidali ricevono e riduce la plasticità sinaptica. Il periodo critico si apre quando i neuroni di Parvalbumin raggiungono la maturazione funzionale; si chiude quando le PNN si consolidano attorno a loro.

La formalizzazione di questo meccanismo in termini di stabilità dinamica è la seguente. Sia $E(t)$ il livello di eccitazione corticale, $I(t)$ il

livello di inibizione (correlato con la maturazione dei neuroni di Parvalbumin), e $P(t)$ il grado di consolidazione delle PNN (che varia da 0 = periodo critico aperto a 1 = periodo critico chiuso). Il sistema dinamico è:

$$dE/dt = f(E, I, \text{Stimolo}) \quad dI/dt = g(E, I, \text{BDNF}) \quad dP/dt = h(I, \text{Tempo}, \text{Esperienza})$$

dove f cattura la dinamica eccitatoria-inibitoria, g l'effetto del BDNF sulla maturazione dei neuroni di Parvalbumin, e h la consolidazione delle PNN in funzione del livello di inibizione, del tempo e dell'esperienza.

Il punto critico del sistema è la biforcazione in $P(t)$: per $P < P_{\text{critico}}$ il sistema è plastico (piccole perturbazioni dei pesi sinaptici si amplificano), per $P > P_{\text{critico}}$ il sistema è stabile (le perturbazioni si estinguono). La chiusura del periodo critico è la transizione da un regime plastico a un regime stabile — una biforcazione di Hopf o di pitchfork nella dinamica della plasticità sinaptica.

III. Path dependence neurale: la struttura formale dell'irreversibilità

III.1 Il concetto di path dependence nei sistemi dinamici

La path dependence — la dipendenza dell'esito finale dalla traiettoria percorsa, non soltanto dalle condizioni attuali — è un fenomeno formalmente preciso nei sistemi dinamici non lineari. Un sistema dinamico è path-dependent se l'insieme dei suoi possibili stati finali dipende dalla sequenza degli stati intermedi, e non soltanto dalle condizioni iniziali e finali.

Formalmente, un sistema dinamico

$$dx/dt = f(x, u(t), t)$$

dove $x \in \mathbb{R}^n$ è il vettore di stato, $u(t) \in \mathbb{R}^m$ è il vettore di controllo esterno (l'esperienza ambientale), e f è il campo vettoriale, è path-dependent se esistono due traiettorie $u_1(t)$ e $u_2(t)$ con la stessa condizione iniziale $x(0)$ e la stessa distribuzione marginale (gli stessi ingressi totali medi, $\int_0^T u_1(t)dt = \int_0^T u_2(t)dt$) che producono stati finali diversi:

$$x_1(T) \neq x_2(T)$$

dove $x_1(T)$ e $x_2(T)$ sono gli stati finali lungo le due traiettorie. La path dependence misura quanto la struttura temporale dell'esperienza (quando avvengono certi eventi) conta, indipendentemente dall'ammontare totale dell'esperienza.

Nei sistemi con periodi critici, la path dependence è massimale: lo stesso ammontare totale di esperienza visiva produce risultati radicalmente diversi a seconda che avvenga durante o dopo il periodo critico. Formalmente, se u_{crit} è l'esperienza durante il periodo critico $[t_{\text{crit}}, t_{\text{crit}} + \Delta t]$ e u_{adult} è la stessa esperienza dopo il periodo critico:

$$\|x(T; u_{\text{crit}}) - x(T; u_{\text{adult}})\| \gg 0$$

anche per $\|u_{\text{crit}} - u_{\text{adult}}\| \approx 0$ (le stesse esperienze prodotte nello stesso modo).

III.2 La termodinamica dell'irreversibilità sviluppativa

L'irreversibilità dei periodi critici può essere formalizzata attraverso un'analogia con la termodinamica fuori dall'equilibrio. Consideriamo lo spazio dei pesi sinaptici W come uno spazio delle fasi termodinamico, con una "energia libera" $F(W)$ che misura la stabilità della configurazione:

$$F(W) = -\log P(W \mid \text{Esperienza})$$

dove $P(W \mid \text{Esperienza})$ è la distribuzione stazionaria dei pesi sinaptici condizionatamente all'esperienza ricevuta. La dinamica sinaptica è analoga a una particella che diffonde in un paesaggio di energia libera con temperatura T (che rappresenta il livello di "fluidità" sinaptica — alta durante il periodo critico, bassa nell'adulto):

$$dW/dt = -\nabla F(W) + \sqrt{(2T)} \eta(t)$$

dove $\eta(t)$ è il rumore bianco. Questa è la dinamica di Langevin per un sistema in contatto con un bagno termico alla temperatura T .

Durante il periodo critico, T è elevata: il sistema esplora ampiamente il paesaggio energetico e può raggiungere stati di equilibrio globale (bassa energia libera) indipendentemente dalla condizione iniziale. Dopo il periodo critico, $T \rightarrow 0$: il sistema è intrappolato in un minimo locale di F , e il rumore termico non è sufficiente a superare le barriere di energia che separano i diversi minimi.

La velocità di transizione tra minimi, nel regime $T \rightarrow 0$, è data dalla formula di Kramers (Kramers, 1940):

$$k \propto \exp(-\Delta F / T)$$

dove ΔF è l'altezza della barriera energetica. Per T molto piccolo (sistema adulto con plasticità ridotta), la velocità di transizione è esponenzialmente piccola — la riorganizzazione dei pesi sinaptici è praticamente impossibile su scale di tempo biologicamente rilevanti. Il sistema è irreversibilmente intrappolato in un minimo locale.

Il minimo locale in cui il sistema si trova nell'adulto dipende dalla traiettoria percorsa durante il periodo critico — la path dependence è formalizzata come l'intrappolamento in minimi locali diversi a seconda della traiettoria. Sistemi che hanno ricevuto esperienze diverse durante il periodo critico si trovano in minimi locali diversi e hanno quindi proprietà funzionali diverse nell'adulto.

III.3 Il teorema di irreversibilità della plasticità sviluppata

Teorema (irreversibilità della plasticità). Sia un sistema neurale con plasticità sinaptica governata dalla dinamica di Langevin con temperatura $T(t)$ decrescente nel tempo ($T \rightarrow 0$ per $t \rightarrow \infty$). Per qualsiasi funzione $T(t)$ strettamente decrescente con $T(t) \rightarrow 0$, il sistema raggiunge asintoticamente un minimo locale di $F(W)$ che dipende dalla traiettoria dell'esperienza durante il periodo in cui $T(t) > T_{\text{critico}}$. Per $t > t_{\text{chiusura}} = \sup\{t : T(t) > T_{\text{critico}}\}$, il sistema è privo di errori con probabilità $1 - O(\exp(-\Delta F_{\text{min}} / T(t)))$.

Dimostrazione. Sia M l'insieme dei minimi locali di $F(W)$. Per $t \geq t_{\text{chiusura}}$, la probabilità di transizione tra minimi adiacenti è bounded superiormente da

$$P(\text{transizione tra } m_1 \text{ e } m_2 \text{ in } [t, t+\Delta t]) \leq \Delta t \cdot C \cdot \exp(-\Delta F_{\text{min}} / T(t))$$

dove $\Delta F_{\text{min}} = \min_{\{m_1 \neq m_2 \in M\}} (F(\text{collo di bottiglia}) - F(m_1))$ è l'altezza minima delle barriere. Per $T(t) \rightarrow 0$, questa probabilità $\rightarrow 0$ uniformemente in t . Quindi il numero atteso di transizioni dopo t_{chiusura} è

$$E[\text{transizioni per } t > t_{\text{chiusura}}] = \int_{t_{\text{chiusura}}}^{\infty} C \cdot \exp(-\Delta F_{\text{min}} / T(t)) dt$$

che è finito per qualsiasi $T(t) \leq C_1 \cdot \exp(-C_2 t)$ con $C_2 > 0$ (plasticità che decade esponenzialmente) — il che è biologicamente plausibile. La serie converge, il numero atteso di transizioni è finito, e quasi certamente il

sistema smette di transitare tra minimi dopo un tempo finito. Il minimo in cui rimane è determinato dalla traiettoria percorsa fino a t_{chiusura} .
□

IV. Il linguaggio come caso paradigmatico: il periodo critico dell'acquisizione linguistica

IV.1 L'ipotesi del periodo critico di Lenneberg

Eric Lenneberg (1967) propose l'ipotesi del periodo critico per l'acquisizione del linguaggio, basandosi su tre linee di evidenza: casi di bambini selvaggi (feral children) che crescono senza esposizione al linguaggio umano; pazienti con lesioni cerebrali che causano afasia a diverse età; e dati sull'acquisizione del linguaggio in bambini sordi esposti a un linguaggio dei segni in tempi diversi.

L'evidenza più drammatica è quella dei bambini selvaggi. Il caso di Genie, una bambina californiana tenuta in isolamento quasi totale dal quarto anno di vita fino ai tredici anni (Curtiss, 1977), mostra che l'esposizione al linguaggio dopo la pubertà produce un'acquisizione parziale: Genie sviluppò un vocabolario considerevole ma non acquisì mai la competenza sintattica — la capacità di costruire e interpretare frasi complesse con struttura gerarchica. L'evidenza anatomica (Curtiss, 1977, 1988) mostra che nel caso di Genie il linguaggio si sviluppò nell'emisfero destro invece del sinistro — il normale substrato neurale del linguaggio — suggerendo una riorganizzazione del substrato neurale conseguente alla mancata attivazione durante il periodo critico.

I dati sull'acquisizione della fonologia mostrano con maggiore precisione la struttura del periodo critico. Werker e Hensch (2015) mostrano che la capacità di distinguere i fonemi non-nativi (i suoni della propria lingua non sono gli stessi di quelli delle altre lingue) è presente nei neonati fino ai sei mesi, si riduce progressivamente fino ai dieci mesi, e non può essere recuperata nell'adulto. Il neonato di ogni cultura distingue tutti i contrasti fonetici di tutte le lingue umane; l'adulto distingue soltanto quelli della propria lingua madre. Questa "sintonizzazione" della percezione fonemica è irreversibile: gli adulti giapponesi non riescono a distinguere il contrasto /r/-/l/ dell'inglese perché quel contrasto non esiste nel giapponese e il loro sistema percettivo fonetico non lo ha codificato durante il periodo critico.

IV.2 La struttura quantitativa del declino della plasticità linguistica

Il declino della plasticità linguistica con l'età segue una funzione di potenza con esponente negativo — non una discontinuità improvvisa (il "periodo critico" nel senso stretto di un interruttore on/off), ma un decadimento progressivo con una struttura che riflette la gerarchia dei sistemi linguistici acquisiti.

Hartshorne, Tenenbaum e Pinker (2018) hanno analizzato i dati di un test di grammatica dell'inglese somministrato online a circa 670.000 partecipanti, ottenendo la più grande analisi mai condotta sulla sensibilità all'età nell'acquisizione del linguaggio. I risultati mostrano che:

La competenza grammaticale degli apprendenti non-nativi in funzione dell'età di immersione nella lingua è descrivibile come:

$$G(A) = G_{\text{max}} \cdot f(A)$$

dove A è l'età di inizio dell'immersione e $f(A)$ è una funzione sigmoideale negativa:

$$f(A) = 1 / (1 + \exp(\beta(A - A_0)))$$

con $A_0 \approx 10\text{--}12$ anni (l'età al centro della curva) e $\beta \approx 0.3 \text{ anni}^{-1}$ (la pendenza della transizione). L'immersione prima dei 10 anni produce competenza pressoché nativa; dopo i 18 anni la competenza raggiunta è proporzionalmente molto inferiore.

Crucialmente, il declino non è uguale per tutti i componenti del sistema linguistico:

La fonologia (i suoni) ha un periodo critico precoce (massima plasticità 0–5 anni) e declina rapidamente.

La morfologia (le forme delle parole) ha un periodo critico intermedio (massima plasticità 0–10 anni).

La sintassi (la struttura delle frasi) ha un periodo critico tardivo (massima plasticità 0–16 anni).

Il lessico (il vocabolario) non ha un periodo critico vero: la capacità di apprendere parole nuove rimane elevata per tutta la vita adulta.

Questa struttura gerarchica — i componenti più fondamentali del linguaggio hanno periodi critici più precoci e più rigidi — rispecchia la struttura evolutiva e developmentale del linguaggio: la fonologia è il

livello più elementare e più indipendente dall'esperienza concettuale, mentre il vocabolario dipende direttamente dall'esperienza del mondo e rimane aperto all'aggiornamento.

V. La neuroplasticità adulta: i limiti e la struttura delle eccezioni

V.1 Il paradosso della neurogenesi adulta

La dogma classica della neuroscienza — "nei adulti il cervello non produce neuroni nuovi" (Ramón y Cajal, 1913) — è stato rovesciato dalla scoperta della neurogenesi adulta nell'ippocampo e nel bulbo olfattivo. Altman (1962) aveva osservato per primo la divisione cellulare nell'ippocampo adulto del ratto, ma la scoperta non fu accettata fino a quando Gould, Reeves, Graziano e Gross (1999) non la confermarono nei primati non umani, e Eriksson et al. (1998) negli esseri umani.

La neurogenesi adulta nell'ippocampo avviene nella zona subgranulare del giro dentato: le cellule staminali neurali si dividono, differenziano in neuroni granulari immaturi, migrano nel giro dentato, e si integrano nei circuiti ippocampali nel corso di settimane. I neuroni nuovi hanno proprietà di plasticità superiori ai neuroni maturi: il loro raggio di eccitazione sinaptica (la curva di risposta in funzione dell'intensità dello stimolo) è più ampio, il potenziale di potenziamento a lungo termine (LTP) è più facilmente indotto, e le soglie di eccitazione sono più basse.

Questo solleva una domanda formale: la neurogenesi adulta non è in contraddizione con la chiusura dei periodi critici? La risposta è che la neurogenesi adulta è una forma limitata e localizzata di plasticità — non una riapertura generalizzata dei periodi critici. La neurogenesi avviene soltanto nell'ippocampo e nel bulbo olfattivo (le due strutture neurali più direttamente legate all'apprendimento episodico e alla discriminazione olfattiva), non nella corteccia sensoriale o nella corteccia prefrontale. La struttura topografica e la specializzazione funzionale delle aree corticali — le "mappe" che sono il prodotto dei periodi critici — rimane stabile nell'adulto.

La formalizzazione di questa limitatezza è la seguente. Sia $\Delta W(t)$ la variazione dei pesi sinaptici in una finestra temporale Δt , e sia W l'insieme dei connettomi delle diverse aree cerebrali. La plasticità residua adulta è:

$$\Delta W(t)_{\text{adulto}} = \Delta W_{\text{hippocampal}} + \Delta W_{\text{olfactory}} + \varepsilon(t) \cdot \Delta W_{\text{cortical}}$$

dove $\epsilon(t) \rightarrow 0$ nell'adulto sano. La neurogenesi adulta modifica soltanto le prime due componenti; la plasticità corticale residua è un termine trascurabile nell'adulto rispetto al periodo critico.

V.2 Il ripristino sperimentale della plasticità: le eccezioni significative

Ricerche recenti hanno dimostrato che la chiusura dei periodi critici non è assoluta: è possibile riaprire sperimentalmente finestre di plasticità nell'adulto attraverso manipolazioni molecolari specifiche. Questi esperimenti, pur non cambiando la conclusione fondamentale sull'irreversibilità nella condizione naturale, rivelano la struttura molecolare della chiusura e aprono prospettive terapeutiche.

Hensch e colleghi hanno dimostrato che la somministrazione di diazepam (un potenziatore del GABA, il principale neurotrasmettitore inibitorio) può aprire un nuovo periodo critico nella corteccia visiva dei gatti adulti: gatti adulti trattati con diazepam mostrano riorganizzazione delle colonne di dominanza oculare in risposta alla privazione monocolare — un effetto che nell'adulto non trattato non si osserva mai. Il meccanismo è il potenziamento prematuro dell'attività dei neuroni di Parvalbumin — la stessa maturazione che normalmente avviene durante il periodo critico naturale ma che nel cervello adulto è soppressa.

Analogamente, la degradazione enzimatica delle reti pericellulari (PNN) attraverso la condroitinasi ABC — un enzima che degrada i proteoglicani costitutivi delle PNN — ripristina la plasticità sinaptica corticale negli adulti. Pizzorusso et al. (2002) hanno mostrato che la condroitinasi ABC somministrata in V1 dei ratti adulti ambliopici ripristina parzialmente la risposta all'occhio ambliopico — un effetto che non si ottiene senza il trattamento enzimatico.

Questi esperimenti dimostrano che la chiusura del periodo critico è una struttura mantenuta attivamente da meccanismi molecolari specifici (le PNN), non uno stato di irreversibilità assoluta della materia cerebrale. La plasticità è soppressa, non eliminata; può essere riattivata, ma a un costo — la destabilizzazione delle strutture già consolidate. Questo è il compromesso fondamentale: la plasticità e la stabilità sono beni rivali nel sistema neurale adulto.

VI. La critica del periodo critico come concetto: la nozione di periodo sensibile

VI.1 La distinzione tra periodo critico e periodo sensibile

La nozione originale di periodo critico è quella di una finestra rigida con confini netti: prima della finestra, nessun effetto permanente; durante la finestra, massima sensibilità; dopo la finestra, nessun effetto. Questa versione discontinua non corrisponde ai dati biologici: le curve di sensibilità all'età sono progressive, non discontinue, e mostrano una transizione graduale piuttosto che un interruttore binario.

Knudsen (2004) ha proposto la distinzione tra periodi critici in senso stretto (effetti permanenti e irreversibili) e periodi sensibili (periodi di massima sensibilità senza irreversibilità assoluta). La maggior parte dei fenomeni studiati rientra nella categoria dei periodi sensibili: la sensibilità all'esperienza è modulata dalla funzione sigmoideale $F(A)$ descritta sopra, con un massimo nel periodo giovanile e una diminuzione graduale nell'adulto.

La struttura matematica della distinzione è la seguente. Sia $P(t)$ la plasticità sinaptica al tempo t e sia $\Delta W(u, t)$ la variazione dei pesi prodotta dall'esperienza u al tempo t . In un periodo critico rigido:

$$P(t) = P_{\max} \cdot \mathbb{1}[t \in [t_{\text{inizio}}, t_{\text{fine}}]]$$

dove $\mathbb{1}$ è la funzione indicatrice. In un periodo sensibile:

$$P(t) = P_{\max} \cdot \sigma(\alpha(t - t_{\text{inizio}})) \cdot (1 - \sigma(\beta(t - t_{\text{fine}})))$$

dove σ è la funzione sigmoide e α, β controllano la steepness delle transizioni. Il periodo critico rigido è il caso limite $\alpha, \beta \rightarrow \infty$.

Il vantaggio adattativo del periodo sensibile rispetto al periodo critico rigido è la robustezza: una finestra graduale permette di recuperare parzialmente da esperienze precoci avverse, mentre una finestra rigida le rende completamente irrecuperabili. D'altro canto, la finestra graduale richiede un meccanismo attivo di diminuzione progressiva della plasticità — il che implica un costo metabolico continuato durante la transizione.

VI.2 La teoria bayesiana dei periodi sensibili

Knudsen (2004) e, più formalmente, Byers e Thornton (2014) hanno proposto una teoria bayesiana dei periodi sensibili che li interpreta come soluzioni ottimali al problema di quando aggiornare un modello generativo del mondo sulla base di nuove osservazioni.

Consideriamo un agente che mantiene un modello del mondo M con distribuzione a priori $P(M)$ e che riceve osservazioni $O(t)$ nel tempo. L'aggiornamento bayesiano ottimale è:

$$P(M \mid O(1:t)) \propto P(O(1:t) \mid M) \cdot P(M)$$

La plasticità — la velocità con cui l'agente aggiorna il modello in risposta a nuove osservazioni — è proporzionale alla varianza posteriore della stima:

$$\eta(t) \propto \text{Var}[P(M \mid O(1:t))]$$

Nelle prime fasi della vita, le osservazioni sono scarse e la varianza posteriore è alta: il sistema deve essere molto plastico per aggiornare rapidamente un modello poco fondato. Man mano che le osservazioni si accumulano, la varianza posteriore diminuisce — il modello diventa più certo — e l'aggiornamento ottimale richiede plasticità minore (un'osservazione aggiuntiva è poco informativa rispetto a molte osservazioni passate).

In questa formulazione, la chiusura del periodo critico è la conseguenza ottimale dell'accumulo di evidenza: un sistema che ha ricevuto molte osservazioni durante il periodo critico ha un modello del mondo ben fondato con varianza posteriore bassa, e ridurre la plasticità è la risposta ottimale che minimizza l'errore quadratico medio della stima nel lungo periodo.

Formalmente, se le osservazioni sono generate da un modello vero M^* con varianza σ^2_{obs} e il prior ha varianza σ^2_{prior} , la varianza posteriore dopo n osservazioni è:

$$\text{Var}[P(M \mid O(1:n))] = \sigma^2_{\text{obs}} \cdot \sigma^2_{\text{prior}} / (n \cdot \sigma^2_{\text{prior}} + \sigma^2_{\text{obs}})$$

Per $n \rightarrow \infty$, la varianza posteriore $\rightarrow 0$ come σ^2_{obs}/n . La plasticità ottimale decresce come $1/n$ — una funzione di potenza della quantità di esperienza accumulata. Questo è formalmente identico alla curva di oblio di Ebbinghaus (discussa nella Monografia III) con la stessa giustificazione: in entrambi i casi, il decadimento come funzione di potenza della quantità di esperienza accumulata è la risposta ottimale bayesiana alla struttura statistica delle esperienze naturali.

VII. Il periodo critico del linguaggio dei segni e la modularità

VII.1 La ricerca su Newport, Supalla e collaboratori

I bambini sordi esposti al linguaggio dei segni della comunità dei sordi (ASL, American Sign Language; LSF, Langue des Signes Française) costituiscono una delle popolazioni sperimentali più preziose per lo studio dei periodi critici, perché permettono di separare l'esposizione al linguaggio dall'esposizione al suono — la variabile che normalmente covaria nei bambini udenti.

Newport, Supalla e collaboratori hanno confrontato la competenza sintattica in ASL di tre gruppi di bambini sordi: nativi (esposti ad ASL dalla nascita, da genitori sordi), precoci (esposti ad ASL da 4 a 6 anni) e tardivi (esposti ad ASL dopo i 12 anni). I risultati sono chiari: i nativi mostrano competenza grammaticale completa, i precoci mostrano errori sistematici in strutture grammaticali complesse (morfologia verbale di accordo, ordine degli argomenti), i tardivi mostrano difficoltà ancora più marcate. Crucialmente, tutti e tre i gruppi mostrano la stessa competenza lessicale — il vocabolario è uguale indipendentemente dall'età di esposizione.

Questo risultato dimostra che il periodo critico del linguaggio dei segni ha la stessa struttura gerarchica del linguaggio parlato: fonologia e morfologia hanno periodi critici più precoci e rigidi rispetto alla sintassi, che a sua volta ha un periodo critico più precoce rispetto al lessico. La modalità del linguaggio (sonora vs. visuo-gestuale) non cambia la struttura del periodo critico — questa è determinata dalla struttura del sistema linguistico stesso, non dall'input sensoriale che lo veicola.

La conseguenza formale è importante: il periodo critico del linguaggio non è un periodo critico della corteccia uditiva (o visiva per i segni), bensì un periodo critico del sistema linguistico in quanto tale — un sistema con struttura computazionale propria che richiede un certo tipo di input durante un certo periodo per strutturarsi completamente. Questo sistema è realizzato in aree corticali diverse a seconda della modalità (temporale-sinistra per il linguaggio parlato, visivo-sinistra per il linguaggio dei segni), ma il periodo critico ha la stessa struttura in entrambi i casi, il che indica che il vincolo è a livello computazionale (cosa viene appreso), non implementazionale (su quale substrato).

VIII. L'imprinting e la path dependence nelle decisioni evolutivamente fondamentali

VIII.1 La struttura dell'imprinting di Lorenz

Konrad Lorenz descrisse il fenomeno dell'imprinting come forma prototipica di apprendimento a periodo critico: i pulcini di anatra, nelle prime ore dopo la schiusa, acquisiscono un'immagine permanente della "madre" identificando qualsiasi oggetto mobile di dimensioni appropriate che si trovi nel loro campo visivo. Questa acquisizione è: (a) limitata a un brevissimo periodo critico (le prime 13–16 ore dopo la schiusa per l'anitra domestica), (b) irreversibile (un'anitra improntata su un essere umano non può essere re-improntata su un'altra anitra), e (c) dotata di conseguenze comportamentali durature che si estendono alle preferenze sessuali nell'adulto.

La struttura matematica dell'imprinting è quella di un problema a di classificazione one-shot: il sistema deve apprendere la categoria "madre" da un singolo esempio (o pochissimi esempi) durante il periodo critico, e quella classificazione diventa permanente. Il meccanismo biologico è la sintesi di proteine che consolidano le sinapsi attivate durante la prima esposizione — un processo che ha la struttura della long-term potentiation (LTP) ma con un meccanismo di consolidazione che è irreversibile senza la disponibilità di certi segnali molecolari (NMDA, CaMKII, CREB) che declinano dopo il periodo critico.

La formalizzazione è la seguente. Sia $C(\text{stimolo})$ la classificazione dello stimolo come "madre" o "non-madre". Durante il periodo critico, il sistema impara:

$$C(\text{stimolo_crit}) = 1 \text{ (madre)} \quad C(\text{qualsiasi altro stimolo}) = 0$$

In modo one-shot, con generalizzazione mediata dalla distanza:

$$C(x) = 1 \text{ sse } d(x, \text{stimolo_crit}) < d_{\text{soglia}}$$

dove d è una metrica nello spazio degli stimoli. Dopo il periodo critico, il classificatore è fisso — non vengono accettati aggiornamenti — qualunque sia l'evidenza nuova.

Il paradosso apparente è che l'imprinting mira all'oggetto "errato" (un essere umano, un palloncino, un'anitra di legno) se quello è il primo oggetto appropriato presentato. La soluzione evolutiva a questo potenziale problema è che nell'ambiente evolutivo di adattamento (un nido naturale), il primo oggetto mobile di dimensioni appropriate che si incontra dopo la schiusa è quasi certamente la madre biologica. L'imprinting è quindi una soluzione bayesiana ottimale all'ambiente naturale — e produce risultati aberranti soltanto in ambienti artificiali (laboratori, allevamenti) che non corrispondono all'EEA.

IX. Conseguenze per la comprensione dei sistemi artificiali

IX.1 L'apprendimento profondo come pseudo-periodo critico

I sistemi di apprendimento profondo (deep learning) mostrano un fenomeno che ha struttura analoga ai periodi critici biologici: la sensibilità ai dati di training è massima nelle fasi precoci dell'addestramento e decresce progressivamente. Il "batch size" ottimale cresce con il progredire dell'addestramento (Smith et al., 2017) — un'osservazione che riflette la diminuzione della varianza posteriore bayesiana man mano che i dati si accumulano. I pesi neurali si "cristallizzano" in una struttura che diventa sempre meno modificabile dai nuovi dati — il fenomeno del catastrophic forgetting, dove l'addestramento su nuovi compiti degrada le prestazioni sui compiti precedenti.

La differenza fondamentale rispetto ai periodi critici biologici è che i sistemi artificiali non hanno un meccanismo molecolare che "chiuda" la plasticità in modo irreversibile: la temperatura di Langevin rimane costante (il tasso di apprendimento η è fisso o decresce lentamente), non collassa a zero come nell'adulto biologico. La path dependence nei sistemi artificiali è quindi meno marcata: con sufficiente addestramento supplementare (fine-tuning), è possibile modificare le rappresentazioni apprese — mentre nel cervello adulto questo è impossibile per le strutture consolidate durante i periodi critici.

Questa differenza strutturale ha conseguenze per la sicurezza dei sistemi AI: le reti artificiali non hanno la "stabilità imposta dalla chiusura del periodo critico" che caratterizza i cervelli biologici. I loro "valori" e le loro "credenze" rimangono potenzialmente modificabili dall'esperienza a qualsiasi età, il che è simultaneamente un vantaggio (plasticità, adattabilità) e un rischio (vulnerabilità alla manipolazione).

X. Conclusioni: il periodo critico come soluzione al trade-off plasticità-stabilità

Il percorso di questa monografia ha rivelato la struttura fondamentale del periodo critico come soluzione evolutiva a un dilemma strutturale: la

plasticità e la stabilità sono beni rivali in qualsiasi sistema di apprendimento fisicamente realizzato. Un sistema completamente plastico — che può essere modificato da qualsiasi esperienza in qualsiasi momento — è instabile: le strutture costruite in risposta all'ambiente precoce vengono continuamente sovrascritte dall'esperienza successiva, producendo il catastrophic forgetting biologico. Un sistema completamente stabile — che non può essere modificato dall'esperienza dopo la nascita — non può adattarsi all'ambiente specifico in cui si trova a vivere, producendo sistemi perfettamente ottimizzati per l'EEA statistico ma non per l'ambiente individuale specifico.

La soluzione evolutiva — una plasticità alta nelle prime fasi della vita che decresce progressivamente fino a un livello residuo basso nell'adulto — è il compromesso ottimale tra queste due esigenze incompatibili. La path dependence che ne risulta — il fatto che lo sviluppo del sistema nervoso dipende irreversibilmente dalla sequenza delle esperienze precoci — non è un difetto evolutivo ma una proprietà necessaria di qualsiasi sistema che risolva questo trade-off.

Il divario tra la potenzialità genetica del cervello umano — le capacità che un cervello con quel genoma potrebbe sviluppare in condizioni ottimali — e le capacità che il cervello individuale effettivamente sviluppa — dipende in modo critico dall'ambiente precoce. Questo è il fondamento biologico dell'importanza della nutrizione precoce, dell'esposizione linguistica precoce, del contatto sociale precoce: non sono lussi educativi, ma presupposti strutturali per la costruzione di un cervello funzionante. La finestra è aperta una sola volta; una volta chiusa, rimane chiusa.

Monografia VI

I limiti della cognizione individuale

Bias come prior mal calibrati, l'ecologia degli errori sistematici, le architetture della razionalità imperfetta

I. L'elogio paradossale dell'errore

Il catalogo dei bias cognitivi umani — compilato nel corso di mezzo secolo da Amos Tversky, Daniel Kahneman e centinaia di collaboratori — ha l'aspetto superficiale di un atto di accusa contro la razionalità

umana. L'uomo è irrazionale: confonde probabilità e frequenze, è ancorato alle prime informazioni ricevute, valuta le perdite più delle vincite equivalenti, ignora i tassi base, proietta il passato nel futuro, sopravvaluta la propria competenza. Questi sembrano difetti da correggere, limitazioni da superare, residui di una preistoria evolutiva che la modernità ha reso obsoleti.

Questa lettura, però, è formalmente scorretta in un senso preciso che questa monografia si propone di dimostrare: i bias cognitivi non sono difetti casuali del sistema cognitivo umano, ma sono la conseguenza necessaria di prior ottimali per l'ambiente evolutivo di adattamento applicati a un ambiente moderno con caratteristiche statistiche radicalmente diverse. Detto altrimenti, nel contesto ecologico appropriato, quelli che chiamiamo bias sono spesso inferenze bayesiane ottimali. La loro riqualificazione come "errori" è un giudizio relativo all'ambiente di test — un ambiente che differisce dall'EEA in dimensioni critiche.

Questo non significa che i bias siano inoffensivi nell'ambiente moderno — producono errori sistematici con conseguenze gravi in medicina, finanza, diritto, politica. Significa che la struttura dei bias non è arbitraria ma riflette la struttura dell'ambiente evolutivo in cui il cervello si è formato, e che questa struttura può essere compresa con la precisione formale della statistica bayesiana e della teoria delle decisioni. Comprendere la struttura è il prerequisito per comprendere quando i bias producono errori nell'ambiente moderno e perché quegli errori non sono facilmente correggibili attraverso l'educazione e la volontà.

II. La struttura bayesiana dei bias cognitivi

II.1 La rappresentatività come prior probabilistico

Il bias di rappresentatività — l'uso della somiglianza a un prototipo come surrogato della probabilità — è forse il più studiato nella letteratura kahneman-tverskyiana. Il classico problema di Linda (Tversky e Kahneman, 1983) mostra che la maggioranza dei soggetti giudica più probabile "Linda è cassiera di banca e attivista femminista" di "Linda è cassiera di banca" — una violazione della monotonia della probabilità. La "fallacy of conjunction" è il nome dell'errore.

La spiegazione bayesiana è la seguente. Il sistema cognitivo non stima $P(\text{cassiera} \mid \text{descrizione})$ e $P(\text{cassiera AND attivista} \mid \text{descrizione})$

attraverso la regola di Bayes con prior statistici. Stima invece la "rappresentatività" — la somiglianza della descrizione al prototipo di ciascuna categoria — e usa questa come stima della probabilità.

Sia d la descrizione di Linda e siano $C_1 = \{\text{cassiera}\}$ e $C_2 = \{\text{cassiera AND attivista}\}$ le due categorie. La probabilità bayesiana corretta è:

$$P(C_2 | d) \leq P(C_1 | d)$$

ma la stima basata sulla rappresentatività è:

$$R(C_2 | d) = \text{sim}(d, \text{prototipo}(C_2)) \quad R(C_1 | d) = \text{sim}(d, \text{prototipo}(C_1))$$

con $R(C_2 | d) > R(C_1 | d)$ perché la descrizione di Linda somiglia di più al prototipo della cassiera-attivista che al prototipo della sola cassiera.

La domanda formale è: quando è R un buon surrogato di P ? Quando la somiglianza al prototipo è una buona stima della probabilità? La risposta bayesiana è: quando i prior su C_1 e C_2 sono proporzionali alla verosimiglianza della descrizione date le categorie:

$$P(C | d) \propto P(d | C) \cdot P(C) \quad P(C | d) \approx R(C | d) \text{ sse } P(C) \approx \text{costante per tutte le } C$$

Se i prior sono uniformi — se tutte le categorie sono ugualmente probabili a priori — allora la somiglianza al prototipo è una buona stima della probabilità. Il prior uniforme è la struttura del tasso base nell'EEA: nei contesti sociali quotidiani, dove gli scopi pratici sono identificare il tipo di persona per prevedere il comportamento, il tasso base delle diverse categorie sociali è relativamente uniforme (non vi è un'enorme disparità tra il numero di cassieri e di cassieri-attivisti nel gruppo di persone che si incontrano quotidianamente). Il bias di rappresentatività produce errori nell'ambiente sperimentale di Tversky e Kahneman perché le categorie hanno tassi base molto diversi (i cassieri sono molto più comuni dei cassieri-attivisti), ma questo tasso base è irrilevante nella pratica quotidiana delle classificazioni sociali.

II.2 La disponibilità come stima di frequenza adattiva

Il bias di disponibilità — la tendenza a giudicare la probabilità di un evento in base alla facilità con cui esempi di quell'evento vengono alla mente — produce errori sistematici nelle stime di frequenza: i voli aerei sono percepiti come più pericolosi degli incidenti stradali (perché gli incidenti aerei sono più coperti dai media), le vittime da squalo come più comuni delle morti da naufragio (stessa ragione), la criminalità come in aumento anche quando decresce (perché i media ne parlano di più nei

periodi di decremento per effetto della "notiziabilità" degli eventi inaspettati).

Ma la disponibilità è una buona stima della frequenza nell'ambiente evolutivo per le classi di eventi che sono sperimentati direttamente o trasmessi attraverso il gossip sociale — non attraverso i media. Nell'EEA, la frequenza con cui esempi di un tipo di evento vengono alla mente è correlata con la frequenza reale di quell'evento nel proprio ambiente immediato: le morti per predazione, le carestie, le alluvioni, le vittorie nella caccia — eventi frequenti sono ricordati di più, eventi rari di meno. La disponibilità è un proxy eccellente per la frequenza quando il campione di informazioni a cui si ha accesso è un campione non distorto dell'ambiente.

I media moderni introducono un bias sistematico in questo campione: segnalano selettivamente gli eventi inusuali (quelli con alta notiziabilità = bassa frequenza), creando una distribuzione dell'informazione accessibile radicalmente diversa dalla distribuzione delle frequenze reali. Il bias di disponibilità non ha subito un aggiornamento evolutivo per adattarsi all'informazione mediata — e non poteva farlo, perché i media sono un'invenzione recentissima rispetto alle scale temporali evolutive.

La struttura formale è precisa. Sia $f(E)$ la frequenza reale dell'evento E nell'ambiente, e sia $m(E)$ la frequenza con cui E è menzionato nei media. La disponibilità $D(E)$ è proporzionale a $m(E)$:

$$D(E) \propto m(E)$$

Il bias produce $P'(E) \propto D(E)$ invece di $P'(E) \propto f(E)$. L'errore è:

$$\epsilon(E) = P'(E) - P(E) \propto m(E) - f(E)$$

Negli ambienti senza media (EEA), $m(E) \approx f(E)$ e l'errore è piccolo. Nell'ambiente moderno con media, $m(E) \neq f(E)$ sistematicamente — la correlazione tra notiziabilità e frequenza è negativa per molte categorie di eventi (gli eventi rari sono più notiziabili di quelli comuni), e l'errore ha segno sistematico.

II.3 L'ancoraggio come aggiornamento bayesiano a partire da un prior

Il bias di ancoraggio — la tendenza a produrre stime insufficientemente aggiornate rispetto a un valore iniziale (l'ancora) — è uno dei bias più robusti e più difficili da eliminare. Tversky e Kahneman (1974) mostrano che anche quando l'ancora è chiaramente casuale (una ruota della fortuna che produce un numero casuale), i soggetti producono stime di quantità sconosciute (la percentuale di paesi africani alle

Nazioni Unite) che sono significativamente influenzate dal numero casuale.

La spiegazione bayesiana dell'ancoraggio interpreta l'ancora come un prior. In un contesto di inferenza bayesiana con prior gaussiano:

$$P(\mu) = N(a, \sigma^2_a)$$

e dati con verosimiglianza gaussiana:

$$P(x | \mu) = N(\mu, \sigma^2_x)$$

la stima posteriore è

$$\hat{\mu}_{\text{MAP}} = (a/\sigma^2_a + x/\sigma^2_x) / (1/\sigma^2_a + 1/\sigma^2_x)$$

che è una media ponderata dell'ancora a e del dato x , con pesi proporzionali alle precisioni (inverse delle varianze) di ciascuna sorgente. L'ancora influenza la stima in proporzione alla sua precisione relativa rispetto ai dati.

Il bias di ancoraggio si manifesta quando σ^2_a è sottostimato rispetto a σ^2_x — ossia quando il sistema cognitivo attribuisce al prior (all'ancora) più peso di quanto sia giustificato dall'evidenza. Perché il sistema cognitivo sottostima la varianza del prior?

Nell'EEA, i valori iniziali che si incontrano in un contesto decisionale sono raramente casuali: sono quasi sempre correlati con il valore vero della quantità da stimare. Il prezzo inizialmente offerto in una trattativa riflette il valore di mercato dell'oggetto; la stima iniziale di una dimensione riflette esperienze passate con dimensioni simili; il numero di oggetti in un gruppo che si vede per primo riflette la distribuzione reale delle grandezze dei gruppi. In questo contesto, usare il valore iniziale come prior informativo è la risposta bayesiana ottimale — non un errore.

Nell'ambiente di Tversky e Kahneman, l'ancora è deliberatamente casuale — completamente non informativa. Ma il sistema cognitivo non dispone di un meccanismo per rilevare l'assenza di correlazione tra l'ancora e il valore vero: per costruzione evolutiva, tratta ogni ancora come potenzialmente informativa. Il risultato è che anche un'ancora casuale viene trattata come un prior con varianza finita, producendo stime parzialmente ancorate.

III. La teoria prospettica e la struttura della valutazione sotto rischio

III.1 La funzione di valore e la curva di avversione alle perdite

Kahneman e Tversky (1979) proposero la teoria prospettica come alternativa descrittiva alla teoria dell'utilità attesa di von Neumann e Morgenstern. La teoria prospettica è costruita su quattro componenti empiricamente fondate: la funzione di valore, la funzione di ponderazione delle probabilità, il punto di riferimento e l'avversione alle perdite.

La funzione di valore $V(x)$ è definita rispetto a un punto di riferimento r (tipicamente lo status quo) e ha la forma:

$$V(x) = \begin{cases} (x - r)^\alpha & \text{se } x \geq r \text{ (guadagni)} \\ -\lambda(r - x)^\beta & \text{se } x < r \text{ (perdite)} \end{cases}$$

con $\alpha, \beta \in (0, 1)$ (concavità per i guadagni, convessità per le perdite — l'effetto di diminishing sensitivity) e $\lambda > 1$ (avversione alle perdite — le perdite pesano più delle vincite equivalenti). I valori empirici tipici sono $\alpha \approx \beta \approx 0.88$ e $\lambda \approx 2.25$ (Tversky e Kahneman, 1992).

La funzione di ponderazione delle probabilità $\pi(p)$ trasforma le probabilità oggettive p in pesi decisionali soggettivi:

$$\pi(p) = p^\gamma / (p^\gamma + (1-p)^\gamma)^{1/\gamma}$$

con $\gamma \approx 0.65$. Questa funzione sovrappesa le probabilità basse ($p \rightarrow 0$: $\pi(p) > p$) e sottopesa le probabilità alte ($p \rightarrow 1$: $\pi(p) < p$). Il valore complessivo di una lotteria $(x_1, p_1; x_2, p_2; \dots; x, p)$ nella teoria prospettica è:

$$V(\text{lotteria}) = \sum_i \pi(p_i) \cdot V(x_i)$$

III.2 La giustificazione evolutiva dell'avversione alle perdite

L'avversione alle perdite ($\lambda > 1$ nella funzione di valore) è spesso citata come l'esempio paradigmatico di irrazionalità: perché la perdita di €100 dovrebbe pesare più del guadagno di €100? Nell'ambiente di un individuo razionale nel senso di von Neumann-Morgenstern, entrambe variano la ricchezza di €100 in direzioni opposte e dovrebbero pesare ugualmente.

Ma la giustificazione evolutiva dell'avversione alle perdite non richiede calcoli di utilità cardinale: richiede soltanto di considerare la distribuzione delle risorse nell'EEA. Kahneman (2011) stesso ha suggerito che l'avversione alle perdite possa essere una risposta adattiva a un'ambiente in cui la perdita di risorse vitali (cibo, territorio, alleati) è più catastrofica per la sopravvivenza della mancata acquisizione di un'uguale quantità di risorse aggiuntive. Se la funzione di benessere biologico è concava nelle risorse (diminishing marginal utility nel gergo economico) con un punto di non-ritorno sotto una soglia minima, allora l'avversione alle perdite è una manifestazione della curvatura di questa funzione attorno allo status quo.

Formalmente, sia $w(R)$ il benessere biologico in funzione delle risorse R . Se w è concava e se R_0 è il livello corrente di risorse, allora per una perturbazione ΔR :

$$|w(R_0 - \Delta R) - w(R_0)| > |w(R_0 + \Delta R) - w(R_0)|$$

per qualsiasi $\Delta R > 0$ e qualsiasi funzione concava w — perché la pendenza di w è maggiore a sinistra di R_0 che a destra. Il valore in termini di benessere biologico di evitare una perdita di ΔR è maggiore del valore di acquisire ΔR . L'avversione alle perdite è la risposta ottimale alla curvatura concava della funzione di benessere biologico.

Il "bias" nasce quando questo meccanismo viene applicato a risorse che non hanno un punto di non-ritorno biologico (il denaro in un'economia moderna) o a decisioni in cui lo status quo non è il punto di riferimento naturale. In questi contesti, la curvatura locale della funzione di valore attorno allo status quo riflette la curvatura della funzione di benessere biologico nell'EEA — non la struttura dell'ambiente moderno.

III.3 La funzione di ponderazione delle probabilità: la sovrastima degli eventi rari

La sovrastima delle probabilità basse ($\pi(p) > p$ per p piccolo) è formalizzata nella funzione di ponderazione delle probabilità della teoria prospettica e produce due effetti apparentemente paradossali: la propensione al rischio per le perdite con bassa probabilità (si preferisce una piccola probabilità di grande perdita a una grande probabilità di piccola perdita equivalente) e la scelta del gioco d'azzardo (si pagano prezzi sopra il valore atteso per basse probabilità di grandi vincite).

La giustificazione evolutiva della sovrastima degli eventi rari è diversa da quella dell'avversione alle perdite: è una questione di segnalazione e di vigilanza. Nell'EEA, gli eventi rari ma catastrofici (incontro con un predatore, collasso della struttura sociale del gruppo, avvelenamento da cibo) richiedono un sistema di allerta che si attivi anche in risposta a

segnali deboli — perché il costo di un falso positivo (scappare da un predatore inesistente) è molto inferiore al costo di un falso negativo (non scappare da un predatore reale).

Questa struttura è quella della teoria della rilevazione del segnale applicata alla decisione sotto rischio. Se il costo C_{FP} del falso positivo è molto inferiore al costo C_{FN} del falso negativo, il criterio ottimale per l'allarme ha la forma:

$$\text{Allarme sse } LR(\text{segnale}) > C_{FP} / C_{FN}$$

dove LR è il likelihood ratio del segnale. Per $C_{FP} \ll C_{FN}$, la soglia di allarme è molto bassa — il sistema scatta per segnali molto deboli — il che equivale a sovrastimare la probabilità dell'evento pericoloso associato al segnale.

Nel linguaggio della teoria prospettica, la sovrastima delle probabilità basse ($\pi(p) > p$) è la versione cognitiva di questo criterio di allerta basso: il sistema tratta gli eventi rari come se fossero più probabili di quanto siano, per assicurarsi di non mancare un evento catastrofico raro. Questo produce errori sistematici nelle decisioni finanziarie (l'acquisto di assicurazioni a prezzi superiori al valore atteso) e nel gioco d'azzardo — ma è adattivo per la gestione dei rischi biologici rari ma catastrofici.

IV. Il sistema duale e la struttura della razionalità limitata

IV.1 La teoria dei processi duali: formalizzazione

Il framework dei processi duali, popularizzato da Kahneman (2011) come "Sistema 1 e Sistema 2", ha una storia intellettuale più profonda che risale agli anni Settanta e Ottanta con i lavori di Wason (1968), Sloman (1996), Evans (1996, 2008) e Stanovich e West (2000). La distinzione di Evans (2008) tra TASS (The Automatic Set of Systems) e AS (The Analytic System) è la più sistematica e formalmente precisa.

Il Sistema 1 (TASS in Evans) è caratterizzato da:

- Elaborazione automatica, non consapevole, non intenzionale
- Alta velocità (tempi di risposta ≤ 200 ms)
- Elaborazione parallela di molteplici cues
- Modularità: specializzazione per domini specifici
- Inaccessibilità introspettiva: non è possibile "guardare dentro" il processo

- Dipendenza dai prior evolutivi e appresi

Il Sistema 2 (AS in Evans) è caratterizzato da:

- Elaborazione controllata, consapevole, intenzionale
- Bassa velocità (tempi di risposta ≥ 500 ms)
- Elaborazione seriale, a capacità limitata
- Generalità: non specializzato per domini
- Accessibilità introspettiva: il processo è (parzialmente) trasparente
- Dipendenza dalle regole esplicite e dal ragionamento formale

La formalizzazione matematica più precisa dei due sistemi usa il framework dell'ottimizzazione vincolata. Il Sistema 1 risolve il problema:

$$\text{Output_S1} = \operatorname{argmax}_{\{o\}} U_{\text{eco}}(o \mid \text{contesto})$$

dove U_{eco} è l'utilità ecologica — la funzione obiettivo ottimizzata dall'evoluzione per l'EEA. Il Sistema 2 risolve il problema:

$$\text{Output_S2} = \operatorname{argmax}_{\{o\}} U_{\text{form}}(o \mid \text{regole, contesto})$$

dove U_{form} è l'utilità formale — la funzione obiettivo definita dal ragionamento esplicito, dalle regole della logica e della probabilità.

In situazioni in cui $U_{\text{eco}} \approx U_{\text{form}}$ (l'ecologia e la razionalità formale concordano), i due sistemi producono lo stesso output. In situazioni in cui $U_{\text{eco}} \neq U_{\text{form}}$ (l'ambiente di test differisce dall'EEA), i due sistemi divergono, e l'output finale dipende dall'esito della "competizione" tra loro.

IV.2 La legge di Evans e la struttura delle credenze implicite

Evans (2006, 2008) ha proposto la "legge di Evans": il pensiero ipotetico-deduttivo richiede la suppression delle credenze di default del Sistema 1 e la loro sostituzione con ipotesi esplicitamente assunte. Questa operazione — la soppressione delle proprie credenze correnti per ragionare in modo condizionale — è computazionalmente costosa e spesso non viene eseguita completamente.

Il test di selezione di Wason (1966) è l'esempio classico. Vengono mostrate quattro carte con lettere e numeri (E, K, 4, 7) e la regola "se una carta ha una vocale su un lato, ha un numero pari sull'altro". Quali carte bisogna girare per verificare la regola? La risposta corretta (logicamente) è E e 7: bisogna verificare se il lato nascosto di E ha un

numero pari (test della sufficienza) e se il lato nascosto di 7 ha una vocale (test della necessità per il contropositivo: se non pari, allora non vocale). La maggioranza dei soggetti sceglie E e 4 — verifica la conferma invece della falsificazione.

L'analisi bayesiana dell'errore è illuminante. Il sistema cognitivo cerca conferme della regola — carteo con E (verifica che abbia un numero pari) e carte con 4 (verifica che abbiano una vocale) — perché nell'EEA le regole vere hanno alta probabilità di conferma. Ma la ricerca di conferme è adattiva per le regole probabilistiche (quasi tutte le regole nell'EEA sono probabilistiche, non deterministiche), non per le regole logiche deterministiche del test di Wason. Il Sistema 1 tratta la regola logica come se fosse probabilistica e cerca conferme — una strategia ottimale nell'EEA ma sbagliata per le regole logiche.

La struttura formale è la seguente. In un contesto bayesiano, il valore informativo di girare la carta E è:

$$I(E) = H(R) - H(R \mid \text{osservazione su } E)$$

dove R è la regola (vera o falsa) e H è l'entropia. Questo è alto sia per la conferma (E ha un numero pari $\rightarrow$ regola vera) che per la falsificazione (E non ha un numero pari $\rightarrow$ regola falsa). Il valore informativo di girare la carta 4 è:

$$I(4) = H(R) - H(R \mid \text{osservazione su } 4)$$

Questo è basso se la regola è quasi certamente vera (come nell'EEA dove le regole osservate di solito sono vere) — girare una carta con un 4 e trovare una vocale non è molto informativo se già si sa che la regola è probabile. Il valore informativo di girare la carta 7 è:

$$I(7) = H(R) - H(R \mid \text{osservazione su } 7)$$

Questo è alto se la regola è probabilistica (trovare una vocale sotto il 7 falsifica certamente la regola). Ma nell'EEA con regole probabilistiche, la falsificazione da parte di un caso ha valore informativo basso — un'eccezione alla regola probabilistica non la falsifica.

La strategia del Sistema 1 (conferma su E e 4) è ottimale per regole probabilistiche nell'EEA. La strategia logicamente corretta (falsificazione su E e 7) è ottimale per regole deterministiche nel contesto formale. L'errore nel test di Wason non è irrazionalità — è l'applicazione di una strategia ottimale per un contesto (regole probabilistiche) a un contesto diverso (regole deterministiche).

V. La calibrazione delle probabilità soggettive

V.1 La curva di calibrazione e l'overconfidence

Una linea di ricerca distinta dagli studi sui bias esplora la calibrazione delle probabilità soggettive — la corrispondenza tra le probabilità che un soggetto assegna agli eventi e le frequenze reali di quegli eventi. Un soggetto perfettamente calibrato assegna probabilità 0.7 a eventi che si verificano nel 70% dei casi, 0.9 a eventi che si verificano nel 90%, e così via.

La curva di calibrazione si ottiene raggruppando le previsioni per intervalli di probabilità soggettiva e plottando la frequenza reale di occorrenza contro la probabilità soggettiva media dell'intervallo. Un soggetto calibrato produce la diagonale principale nel piano (probabilità soggettiva, frequenza reale); overconfidence produce punti sopra la diagonale per le basse probabilità e sotto la diagonale per le alte probabilità (si afferma troppa certezza); underconfidence produce il pattern opposto.

Il risultato più robusto nella letteratura sulla calibrazione è l'overconfidence per le domande di conoscenza generale: domande del tipo "Qual è la capitale della Turchia? Quanto sei sicuro della risposta (50%–100%)?". I soggetti che affermano di essere al 90% sicuri hanno ragione circa nel 70% dei casi. La curva di calibrazione mostra overconfidence sistematica per le alte probabilità soggettive.

La formalizzazione matematica usa la score di Brier (Brier, 1950) come misura della calibrazione complessiva:

$$BS = (1/n) \sum_i (p_i - o_i)^2$$

dove p_i è la probabilità soggettiva assegnata all'evento i e $o_i \in \{0,1\}$ è l'esito osservato. Il punteggio di Brier è scomponibile in tre componenti:

$$BS = \text{Calibrazione} + \text{Risoluzione} - \text{Incertezza}$$

dove la Calibrazione misura la corrispondenza tra probabilità soggettive e frequenze reali (idealmente zero), la Risoluzione misura la capacità del soggetto di discriminare tra eventi che accadranno e eventi che non accadranno (idealmente alta), e l'Incertezza misura la varianza degli esiti (indipendente dal soggetto).

La giustificazione adattiva dell'overconfidence (nelle aree in cui si manifesta) è legata alla funzione dell'autostima come segnale sociale. Nell'EEA, gli individui che manifestavano elevata fiducia nelle proprie capacità ottenevano più risorse, più alleanze e più opportunità riproduttive — indipendentemente dall'accuratezza di quella fiducia. L'overconfidence è quindi, in parte, un'adattamento sociale-segnalatorio invece che un difetto epistemico.

V.2 La struttura degli errori di commissione e omissione

Un secondo aspetto della calibrazione dei bias è l'asimmetria tra errori di commissione (fare qualcosa di sbagliato) e omissione (non fare qualcosa di giusto). I soggetti tipicamente assegnano maggiore responsabilità morale e maggiore rimpianto agli errori di commissione che a errori di omissione equivalenti in termini di esito.

Il bias di omissione è particolarmente evidente in medicina: i medici sono più propensi a non prescrivere un trattamento potenzialmente rischioso (anche quando statisticamente benefico) che a prescriverlo — perché la responsabilità percepita per un danno causato dal trattamento (commissione) è maggiore della responsabilità percepita per un danno causato dalla malattia non trattata (omissione).

La formalizzazione usa la teoria dell'attribuzione causale. Sia D un danno. Il sistema cognitivo attribuisce responsabilità in proporzione alla causalità percepita:

$R(\text{agente}, D) \propto P(\text{controfattuale: se agente non avesse agito, } D \text{ non sarebbe avvenuto})$

Per un danno causato dall'azione dell'agente (commissione), il controfattuale "se l'agente non avesse agito, D non sarebbe avvenuto" è quasi certamente vero — la responsabilità è alta. Per un danno causato dall'inazione dell'agente (omissione), il controfattuale "se l'agente avesse agito, D non sarebbe avvenuto" è meno certo — dipende dall'efficacia dell'intervento, che è probabilistica. L'asimmetria nella certezza del controfattuale produce l'asimmetria nella responsabilità percepita.

Nell'EEA, questa asimmetria è adattiva: gli errori di commissione sono più visibili socialmente (l'azione produce effetti osservabili immediati), mentre gli errori di omissione sono meno visibili (la mancanza di azione non produce effetti immediati distinguibili dalla normalità). La pressione selettiva sulla responsabilità per le commissioni è stata quindi più forte di quella sulle omissioni, producendo l'asimmetria osservata.

VI. I limiti dell'introspezione: la cecità per il cambiamento e la cecità a se stessi

VI.1 Il change blindness e l'incoerenza delle scelte

L'incapacità di percepire cambiamenti nel campo visivo durante saccadi o interruzioni del segnale visivo — il change blindness (Rensink, O'Regan e Clark, 1997) — è una delle dimostrazioni più drammatiche della natura costruttiva (non registrativa) della percezione visiva. I soggetti che guardano un video in cui oggetti salienti cambiano durante un taglio (interruzione del flusso visivo) spesso non riescono a percepire il cambiamento, anche se l'oggetto cambiato è di dimensioni considerevoli e occupa una posizione centrale nel campo visivo.

Ma il change blindness non è soltanto un fenomeno visivo. Johansson et al. (2005, 2006) hanno dimostrato il "choice blindness": i soggetti che scelgono tra due opzioni (due fotografie, due marmellate, due affermazioni morali) e poi vengono presentati con l'opzione non scelta come se fosse quella scelta tipicamente non rilevano la sostituzione e, quando viene chiesto loro di giustificare la "propria" scelta, producono razionalizzazioni coerenti dell'opzione non scelta.

Questo fenomeno mostra che l'accesso introspettivo ai propri processi decisionali è altamente limitato: il sistema cognitivo non ha accesso diretto alle cause delle proprie scelte, ma costruisce spiegazioni post-hoc che sono coerenti con il risultato osservato (l'opzione presentata come scelta) indipendentemente da come la scelta sia stata effettivamente effettuata.

La formalizzazione è quella della teoria dell'accesso alla coscienza. Il Sistema 1 produce una scelta attraverso processi non accessibili alla coscienza. Il Sistema 2 riceve l'output della scelta (ma non il processo che l'ha generata) e costruisce una narrazione esplicativa. La narrazione è plausibile — si adatta all'output — ma non è causalmente accurata rispetto al processo interno. Nel choice blindness, la narrazione si adatta all'output sbagliato (l'opzione non scelta) perché il Sistema 2 non ha accesso alla storia del processo decisionale.

VI.2 L'effetto Dunning-Kruger e la metacognizione difettiva

Kruger e Dunning (1999) hanno documentato un fenomeno sistematico nella metacognizione — la valutazione delle proprie prestazioni: le persone con competenza inferiore tendono a sovrastimare le proprie prestazioni rispetto alla media, mentre le

persone con competenza superiore tendono a sottostimarle. Questo effetto è stato documentato in logica, grammatica, umorismo e competenza medica.

La spiegazione di Kruger e Dunning è che la competenza nel dominio è necessaria non soltanto per la performance ma anche per la metacognizione: per valutare le proprie prestazioni, è necessario avere il modello del dominio che permette di identificare gli errori. Chi manca di competenza non può riconoscere i propri errori — perché non ha il modello che li definirebbe come errori.

La formalizzazione usa la nozione di modello metacognitivo. Sia M il modello cognitivo del dominio (la conoscenza di come ragionare in quel dominio), e sia M' il modello metacognitivo (la conoscenza di come valutare la propria competenza nel dominio). In molti domini, $M' \subset M$ — il modello metacognitivo è un sottoinsieme del modello cognitivo, e non si può avere M' senza M . Di conseguenza:

$$\text{Accuratezza_metacognitiva}(\text{soggetto}) \approx f(\text{Competenza_dominio}(\text{soggetto}))$$

con f monotona crescente. I soggetti a bassa competenza (basso M) hanno basso M' e quindi bassa accuratezza metacognitiva — non riescono a valutare accuratamente le proprie prestazioni. I soggetti ad alta competenza (alto M) hanno alto M' e quindi alta accuratezza metacognitiva.

L'effetto Dunning-Kruger non è una forma di arroganza o di malafede — è la conseguenza necessaria della struttura dipendente del modello metacognitivo dal modello cognitivo. Non si può sapere quello che non si sa.

VII. La funzione adattiva dei bias: quando l'irrazionalità è razionale

VII.1 Il criterio di ottimalità ecologica

La ricerca di Gigerenzen e collaboratori (ABC Research Group) ha sistematicamente dimostrato che in certi contesti naturali, le euristiche "irrazionali" — quelle che violano i canoni della teoria dell'utilità attesa — superano le strategie formalmente ottimali in termini di accuratezza predittiva su dati nuovi, velocità di risposta e robustezza al rumore.

Il risultato più citato è il paradosso della certezza (Gigerenzen, Czerlinski e Martignon, 1999): l'euristica "Take the Best" (descritta nella Monografia III) supera la regressione lineare multipla nella predizione out-of-sample in una serie di problemi reali (previsione degli episodi cardiaci, identificazione dei candidati universitari, previsione delle quote di borsa) quando il numero di campioni di training è limitato. La superiorità è nel regime $N/K < c^*$ dove N è il numero di campioni e K è il numero di predittori, con $c^* \approx 10$ per problemi tipici.

La spiegazione è formale: la regressione multipla è un stimatore non vincolato che stima K parametri dagli N dati, e quando N/K è piccolo la varianza dello stimatore è alta — la regressione overfits. L'euristica Take the Best è uno stimatore fortemente vincolato (usa un solo predittore, il più valido) con varianza molto bassa — al costo di un bias più alto. Il tradeoff bias-varianza favorisce Take the Best quando N è piccolo.

Il parallelismo con l'EEA è preciso: nell'EEA, le decisioni tipiche (scegliere quale via percorrere, scegliere quale alleato fidarsi, stimare la dimensione di una preda) si basano su campioni limitati di informazione (pochi incontri precedenti con situazioni simili), e le euristiche a bassa varianza sono sistematicamente superiori alle strategie ottimali che richiedono campioni grandi per calibrare i propri parametri.

VII.2 Il less-is-more effect e la struttura dell'ottimale ecologico

Il "less-is-more effect" documentato da Gigerenzen — la superiorità delle strategie che usano meno informazione rispetto a quelle che ne usano di più — ha una struttura formale precisa nel trade-off bias-varianza. Sia $MSE(\hat{h}) = Bias^2(\hat{h}) + Var(\hat{h})$ l'errore quadratico medio di uno stimatore $\hat{h}$. La selezione tra stimatori dovrebbe minimizzare il MSE totale, non il bias o la varianza individualmente.

Per le euristiche ecologiche, il bias è strutturalmente non zero (usano informazione incompleta e assumono una struttura specifica dell'ambiente) ma la varianza è molto bassa (pochi parametri da stimare, alta stabilità alle fluttuazioni del campione). Per gli stimatori statistici formali, il bias è basso (usano tutta l'informazione disponibile) ma la varianza è alta se il campione è piccolo.

Il punto di crossover — la dimensione del campione N^* al di sotto della quale l'euristica supera lo stimatore formale — dipende dalla struttura del problema. Per il caso lineare con K predittori e il confronto tra Take the Best e regressione multipla:

$$N^* \approx 2K \text{ (approssimativamente)}$$

Per $N < N^*$, Take the Best ha MSE inferiore; per $N > N^*$, la regressione multipla ha MSE inferiore. Poiché nell'EEA N è tipicamente dell'ordine di poche decine (il numero di situazioni simili che si incontrano in una vita), e K è dell'ordine di poche unità (i predittori rilevanti per una decisione), il rapporto N/K è dell'ordine di 5–20 — esattamente nel regime in cui Take the Best è ottimale.

VIII. I limiti intrinseci della debiasing: perché i bias sono resistenti alla correzione

VIII.1 La resistenza all'istruzione formale

Una delle osservazioni più sconcertanti nella letteratura sui bias cognitivi è che la conoscenza dei bias — sapere che esistono, avere ricevuto istruzione formale in statistica e probabilità — non li elimina. I soggetti con formazione avanzata in statistica continuano a commettere molti degli stessi errori dei soggetti naïf nelle condizioni sperimentali originali. I medici con anni di esperienza nel ragionamento diagnostico continuano a mostrare bias di ancoraggio, di disponibilità e di rappresentatività. I giudici professionisti continuano a mostrare effetti di framing nelle sentenze.

La spiegazione strutturale è la seguente. I bias sono implementati nel Sistema 1 — il sistema automatico e modulare che opera al di sotto della soglia della coscienza. L'istruzione formale agisce principalmente sul Sistema 2 — il sistema controllato e consapevole che può ragionare esplicitamente sui bias. Ma il Sistema 1 opera prima del Sistema 2 e produce l'output iniziale che il Sistema 2 poi (eventualmente) modifica.

La modifica del Sistema 2 è computazionalmente costosa: richiede il riconoscimento che una situazione di bias si sta verificando (il che richiede a sua volta conoscenza del bias e riconoscimento del pattern), la soppressione dell'output del Sistema 1 (che richiede inibizione cognitiva), e la sostituzione con il ragionamento normativo (che richiede elaborazione formale). Questo processo richiede capacità cognitive disponibili — il che significa che avviene soltanto quando il soggetto è "a suo agio" (non sotto pressione di tempo, non cognitivamente sovraccarico, non emotivamente eccitato). In condizioni naturali di pressione temporale e di carico cognitivo elevato — le condizioni in cui molte decisioni importanti devono essere prese — il Sistema 2 non ha le risorse per correggere il Sistema 1.

VIII.2 Il nudging come alternativa alla debiasing

Thaler e Sunstein (2008) hanno proposto l'approccio del nudging come alternativa più efficace alla debiasing diretta: invece di modificare il processo cognitivo (il Sistema 1), si modifica l'architettura delle scelte in modo che il sistema cognitivo naturale — con i suoi bias — produca risultati più vicini a quelli normativamente desiderabili.

Il nudging sfrutta i bias invece di combatterli. Il default bias (la tendenza a mantenere l'opzione di default senza agire attivamente per cambiare) è usato nel design dei piani di previdenza integrativa: rendere l'iscrizione automatica (con opt-out) invece di volontaria (con opt-in) aumenta drasticamente il tasso di partecipazione — sfruttando il bias di inazione e il default effect invece di cercare di eliminarli. Il framing effect è sfruttato nella comunicazione medica: presentare un trattamento come "90% di sopravvivenza" invece di "10% di mortalità" aumenta l'accettazione del trattamento — sfruttando la sensitivity diversa ai guadagni e alle perdite.

La struttura formale del nudging è quella dell'ottimizzazione dell'architettura delle scelte. Data una funzione cognitiva $\varphi : \text{Scelta} \rightarrow \text{Decisione}$ che riflette i bias del Sistema 1, e dato un obiettivo normativo O (la decisione che il soggetto vorrebbe prendere con piena razionalità), il nudge è una trasformazione dello spazio delle scelte $C \rightarrow C'$ tale che:

$$\varphi(C') \approx O$$

Il nudging "inganna" il Sistema 1 modificando l'ambiente in modo che i suoi output coincidano (approssimativamente) con gli output normativamente desiderabili, senza modificare il Sistema 1 stesso.

IX. Conclusioni: i limiti come architettura, non come difetti

Il catalogo dei bias cognitivi umani, esaminato con la strumentazione formale della statistica bayesiana e della teoria delle decisioni, non è la lista dei difetti di una razionalità imperfetta da correggere. È la mappa dell'architettura cognitiva di una specie costruita per prosperare in un ambiente specifico — l'Africa del Pleistocene — con risorse computazionali limitate e pressioni temporali elevate.

I bias di rappresentatività, disponibilità e ancoraggio sono prior bayesiani ottimali per l'EEA applicati a un ambiente con caratteristiche statistiche diverse. L'avversione alle perdite è la risposta ottimale alla

curvatura concava della funzione di benessere biologico attorno allo status quo. La sovrastima delle probabilità basse è il criterio di allerta ottimale per eventi rari e catastrofici. L'overconfidence è un adattamento sociale-segnalatorio. Il change blindness è la conseguenza della natura costruttiva (non registrativa) della percezione. L'effetto Dunning-Kruger è la conseguenza della dipendenza strutturale del modello metacognitivo dal modello cognitivo.

Questa rilettura non è una scusa per i bias — non nega che producano errori sistematici nell'ambiente moderno, spesso con conseguenze gravi. È una comprensione della loro struttura che permette di prevedere esattamente quando producono errori (quando l'ambiente moderno differisce dall'EEA) e che suggerisce strategie di correzione appropriate (nudging, non debiasing). È, in ultima analisi, la dimostrazione che i limiti della cognizione individuale hanno la stessa struttura dei limiti dell'ottimizzazione evolutiva analizzata nella Monografia II: sono la conseguenza necessaria di soluzioni localmente ottimali a problemi generalmente impossibili, costruite in un ambiente specifico che la modernità ha trasformato più rapidamente di quanto l'evoluzione possa seguire.

Monografia VII

I limiti della trasmissione culturale

L'effetto cricchetto di Tomasello, la deriva culturale, la soglia demografica di Henrich e le impossibilità della fedeltà trasmissiva

I. La cultura come sistema di ereditarietà e le sue anomalie

La trasmissione culturale — il passaggio di informazione, competenze, credenze e pratiche da individuo a individuo e da generazione a generazione — è il meccanismo che distingue in modo più netto la cognizione umana da quella degli altri primati. Non soltanto gli esseri umani accumulano conoscenza nel corso della propria vita individuale (apprendimento individuale); non soltanto la trasmettono ai propri figli e discendenti (trasmissione verticale); la trasmettono in modo sufficientemente fedele da permettere che le innovazioni di una generazione si sommino a quelle delle generazioni precedenti,

producendo nel corso del tempo tecnologie, sistemi simbolici, istituzioni e pratiche di complessità crescente.

Michael Tomasello (1999) ha battezzato questo processo "effetto cricchetto" (ratchet effect): la cultura umana avanza per passi, ciascuno dei quali si appoggia sui passi precedenti senza tornare indietro — come un cricchetto che ruota in una sola direzione. L'evoluzione culturale differisce dall'evoluzione biologica non soltanto nella velocità (tempi di generazioni culturali di anni o decenni contro tempi di generazioni biologiche di anni o secoli), ma nella struttura: l'informazione culturale può fluire in qualsiasi direzione nella rete sociale (trasmissione orizzontale tra coetanei, trasmissione obliqua da adulti non-genitori ai giovani), può accumularsi in artefatti e istituzioni che persistono oltre la vita degli individui, e può essere deliberatamente innovata e migliorata.

Eppure anche la trasmissione culturale — per quanto potente rispetto all'evoluzione biologica — incontra limiti strutturali che questa monografia si propone di analizzare con rigore formale. Questi limiti non sono semplicemente limitazioni pratiche (insufficienza dei canali di trasmissione, incapacità cognitiva dei trasmettitori o dei ricevitori), bensì impossibilità strutturali che derivano dalla natura dell'informazione trasmessa, dalla capacità dei sistemi cognitivi di elaborarla e dalla struttura demografica delle popolazioni che la mantengono. Il filo conduttore è la dimostrazione che la fedeltà della trasmissione culturale è condizionata da fattori demografici, cognitivi e informativi che in certi regimi producono degenerazione — non progresso — della complessità culturale.

II. La struttura dell'apprendimento sociale e le sue varietà

II.1 La tassonomia dell'apprendimento sociale

L'apprendimento sociale — l'acquisizione di informazione attraverso l'osservazione o l'interazione con altri individui — è un meccanismo diffuso nel regno animale: topi apprendono rotte, uccelli apprendono canti, primati apprendono tecniche di foraggiamento osservando i conspecifici. Ma l'apprendimento sociale umano ha proprietà qualitativamente distintive che Tomasello ha sistematizzato nella distinzione tra imitazione, emulazione e insegnamento.

L'emulazione è la forma più elementare: l'osservatore impara lo scopo dell'azione (ottenere il cibo da dentro il tubo) e lo raggiunge con mezzi

propri, senza replicare i gesti specifici dell'attore. Gli scimpanzé mostrano prevalentemente emulazione: osservano lo scopo e trovano soluzioni proprie. L'imitazione è l'acquisizione dei mezzi specifici dell'azione — la replicazione dei gesti specifici dell'attore, non soltanto del suo scopo. I bambini umani mostrano imitazione anche quando i gesti "inutili" dell'attore non sono necessari per raggiungere lo scopo (overimitation documentata da Horner e Whiten, 2005). L'insegnamento è la trasmissione intenzionale di informazione da un individuo che la possiede a un individuo che non la possiede, con consapevolezza dello stato epistemico dell'altro (si insegna soltanto quando si sa che l'altro non sa).

La distinzione è formalmente importante perché la fedeltà della trasmissione dipende dal meccanismo. L'emulazione trasmette l'obiettivo funzionale ma non il metodo — l'informazione trasmessa è il problema, non la soluzione, e la soluzione deve essere reinventata in ogni generazione. L'imitazione trasmette sia il metodo che l'obiettivo, permettendo l'accumulo delle innovazioni metodologiche. L'insegnamento trasmette sia il metodo che le ragioni del metodo, permettendo la trasmissione di regole generali invece di casi specifici.

La formalizzazione usa il linguaggio della teoria dell'informazione. Sia θ il parametro della tecnica culturale (l'insieme delle caratteristiche che la definiscono) e sia o la prestazione osservabile dell'esecutore della tecnica. L'informazione trasmessa attraverso i diversi meccanismi è:

Emulazione: $I_{\text{trasmessa}} = I(\theta_{\text{obiettivo}}; o)$ — informazione sull'obiettivo
 Imitazione: $I_{\text{trasmessa}} = I(\theta_{\text{completo}}; o)$ — informazione sul metodo completo
 Insegnamento: $I_{\text{trasmessa}} = I(\theta_{\text{completo}}, \text{regola}; o)$ — metodo e struttura profonda

Per una tecnica con parametro θ di complessità $K(\theta)$ (complessità di Kolmogorov), l'imitazione fedele trasmette quasi tutta l'informazione, ma c'è sempre un residuo di perdita di informazione per osservazione:

$$I_{\text{persa}} = K(\theta) - I(\theta; o) = H(\theta | o) > 0$$

Questa perdita di informazione per ciclo di trasmissione è il seed della deriva culturale — l'equivalente culturale delle mutazioni genetiche.

II.2 La fedeltà della trasmissione e il modello di copia-con-errore

Il modello più semplice della trasmissione culturale con errori — l'analogo culturale della replicazione del DNA — è il modello di copia-con-errore introdotto da Cavalli-Sforza e Feldman (1981) e Boyd e Richerson (1985). In questo modello, un individuo della generazione

$t+1$ apprende da un modello della generazione t copiando la variante culturale c del modello con probabilità $1-\epsilon$ e innovando (adottando una variante diversa casuale) con probabilità ϵ .

La dinamica delle frequenze culturali è analoga alla dinamica genetica: se $f_i(t)$ è la frequenza della variante culturale i nella generazione t , l'aggiornamento è:

$$f_i(t+1) = (1-\epsilon) \cdot f_i(t) + \epsilon \cdot \mu_i$$

dove μ_i è la frequenza di innovazione della variante i . Per ϵ piccolo e μ_i uniforme, il sistema è dominato dalla deriva culturale — la variante dominante è fissata per un tempo medio proporzionale a $1/\epsilon$ (analogo al tempo di fissazione genetica in una popolazione di taglia effettiva $1/\epsilon$).

La differenza cruciale rispetto alla genetica è la multigenitorialità culturale: un individuo può apprendere da molti modelli simultaneamente (insegnanti, pari, figure culturali), e la frequenza efficace di errore è ridotta rispetto alla trasmissione da un singolo modello:

$$\epsilon_{\text{eff}} = \epsilon / n_{\text{modelli}}$$

dove n_{modelli} è il numero di modelli da cui il soggetto apprende. In popolazioni grandi con molti modelli, la deriva culturale è lenta; in popolazioni piccole con pochi modelli, la deriva è rapida e può portare a degenerazione della complessità.

III. La soglia demografica di Henrich: il paradosso della complessità perduta

III.1 Il caso Tasmania e il modello demografico

Joseph Henrich (2004) ha proposto un modello formale per spiegare uno dei paradossi più sorprendenti dell'archeologia preistorica: la perdita progressiva di complessità tecnologica nei popoli della Tasmania durante i 10.000 anni di isolamento che seguirono la separazione dal continente australiano per l'innalzamento del livello del mare circa 11.000 anni fa.

I tasmaniani al momento del primo contatto europeo (fine del XVIII secolo) avevano una delle tecnologie più semplici documentate nella storia umana: nessun osso lavorato, nessuna rete da pesca, nessuna ascia

di pietra levigata, nessun abito di pelle cucito, nessuna tecnica di accensione del fuoco. Eppure gli antenati dei tasmaniani provenivano dall'Australia continentale, dove 40.000 anni fa avevano tecnologie più sofisticate — tecnologie che i tasmaniani, nel corso di 10.000 anni di isolamento, avevano perso.

Il modello di Henrich propone che la dimensione della popolazione — e specificamente la dimensione efficace della rete di apprendimento sociale — sia il fattore determinante della complessità tecnologica sostenibile. Il modello formalizza l'intuizione che la conoscenza complessa richiede una rete sociale abbastanza grande da contenere un numero sufficiente di individui con competenza elevata.

III.2 Il modello formale di Henrich

Sia $\theta \in [0, \theta_{\max}]$ il livello di abilità in una tecnica culturale, dove $\theta_{\max}$ è il livello ottimale. Il modello assume che:

1. Ogni individuo apprende dalla persona più abile tra i propri n modelli sociali.
2. L'apprendimento è soggetto a errore: l'abilità acquisita è $z = \theta_{\text{modello}} + \epsilon$, dove $\epsilon \sim N(0, \sigma^2)$ è il rumore di apprendimento.
3. Il livello di abilità è delimitato superiormente: $z = \min(\theta_{\text{modello}} + \epsilon, \theta_{\max})$.

Sia $Z = \max(X_1, X_2, \dots, X_n)$ il livello dell'apprendice al tempo $t+1$, dove $X_1, \dots, X_n \sim N(\theta, \sigma^2)$ sono i livelli degli n modelli (con media θ e varianza σ^2). Il livello atteso dell'apprendice è:

$$E[Z] = E[\max(X_1, \dots, X_n)] \approx \theta + \sigma \cdot E_{\max}(n)$$

dove $E_{\max}(n) = E[\max(Z_1, \dots, Z_n)]$ con $Z_i \sim N(0,1)$ è il valor atteso del massimo di n variabili normali standard. Per n grande, $E_{\max}(n) \approx \sqrt{2 \ln n}$ per il teorema degli estremi del massimo gaussiano.

La condizione di equilibrio — il livello di abilità medio si mantiene costante attraverso le generazioni — richiede:

$$E[Z] = \theta$$

ovvero:

$$\theta + \sigma \cdot E_{\max}(n) = \theta$$

Questo non può mai essere soddisfatto esattamente a causa del termine $\sigma \cdot E_{\max}(n) > 0$ — il livello atteso dell'apprendice è sempre superiore alla media dei modelli perché apprende dalla persona migliore tra n

modelli. Il termine $\sigma \cdot E_{\max}(n)$ rappresenta il guadagno per generazione dalla selezione del modello migliore.

Ma questo presuppone che l'abilità sia delimitata superiormente da $\theta_{\max}$. All'equilibrio, la distribuzione delle abilità si concentra vicino a $\theta_{\max}$, e il guadagno è bilanciato dalle perdite al bordo superiore (l'abilità non può superare $\theta_{\max}$). L'equilibrio è raggiunto quando:

$$P(z > \theta_{\max} \mid \theta) = \text{probabilità di perdita al bordo}$$

Questo dà una distribuzione stazionaria dell'abilità con media $\theta^*(n)$ che dipende da n . Il punto cruciale è che per n piccolo:

$$E_{\max}(n) \approx \sigma \sqrt{2 \ln n}$$

è piccolo, e il guadagno per generazione dalla selezione del modello migliore è insufficiente a controbilanciare le perdite per errore di apprendimento, specialmente per le tecniche complesse (alto σ). L'abilità media degenera progressivamente.

Henrich deriva la soglia demografica n^* come il valore di n al di sotto del quale la complessità tecnologica non è sostenibile:

$$n^* = \exp((\theta_{\max}/\sigma)^2 / 2)$$

Per $\theta_{\max}/\sigma = 3$ (la complessità della tecnica è tre volte la deviazione standard del rumore di apprendimento), $n^* \approx \exp(4.5) \approx 90$. Una rete di apprendimento con meno di 90 individui non può sostenere una tecnica di questa complessità: la degenerazione è inevitabile nel lungo periodo.

III.3 La verifica del modello: evidenza cross-culturale

Henrich (2004) verifica il modello attraverso la correlazione cross-culturale tra dimensione della rete sociale e complessità tecnologica nelle popolazioni dell'isole della Polinesia e del Pacifico. In queste popolazioni, la dimensione demografica varia di diversi ordini di grandezza (da poche centinaia di individui a decine di migliaia) e la complessità tecnologica è documentata attraverso l'archeologia e l'etnografia.

Il risultato è una correlazione positiva significativa tra dimensione demografica e complessità degli strumenti:

$$C(\text{pop}) \approx a + b \cdot \log(N)$$

dove C è la complessità degli strumenti (misurata come numero di componenti distinte, numero di passaggi produttivi, o indice di sofisticazione materiale), N è la dimensione della rete di apprendimento

e $b > 0$ è il coefficiente di scala. La relazione è logaritmica come previsto dal modello: ogni raddoppio della dimensione della rete produce un incremento additivo della complessità sostenibile, proporzionale a $\sqrt{(2 \ln 2)} \cdot \sigma$.

La deriva demografica — la riduzione della dimensione della rete per isolamento geografico, catastrofi, epidemie — produce invariabilmente riduzione della complessità tecnologica nella finestra temporale di alcune generazioni. La selezione culturale può rallentare questa degenerazione (conservare preferenzialmente le tecnologie più utili), ma non può fermarla quando la rete scende sotto la soglia critica n^* .

IV. La struttura formale della deriva culturale

IV.1 Il modello di Wright-Fisher culturale

La deriva culturale — il cambiamento casuale nelle frequenze culturali per effetto del campionamento nelle popolazioni finite — è formalmente analoga alla deriva genetica. Il modello di Wright-Fisher culturale (Cavalli-Sforza e Feldman, 1981) descrive la dinamica delle frequenze di varianti culturali in una popolazione di N individui.

Sia $f_i(t)$ la frequenza della variante culturale i al tempo t . Se ogni individuo della generazione $t+1$ sceglie casualmente un modello dalla generazione t e adotta la sua variante (trasmissione verticale pura senza selezione), la frequenza della variante i nella generazione $t+1$ è:

$$f_i(t+1) = (1/N) \cdot \sum X$$

dove $X \in \{0,1\}$ è la variabile indicatrice "il j -esimo individuo ha scelto la variante i ", con $P(X=1) = f_i(t)$. La distribuzione di $f_i(t+1)$ data $f_i(t)$ è dunque binomiale:

$$N \cdot f_i(t+1) \mid f_i(t) \sim \text{Binomiale}(N, f_i(t))$$

con media $f_i(t)$ e varianza $f_i(t)(1-f_i(t))/N$.

Questo modello produce la stessa dinamica del modello genetico: in assenza di selezione culturale (tutte le varianti hanno uguale "fitness culturale"), tutte le varianti si fissano o si estinguono per deriva, con probabilità di fissazione uguale alla frequenza iniziale e tempo medio di fissazione pari a $4N$ generazioni culturali.

La differenza fondamentale rispetto alla genetica è il meccanismo di "selezione" culturale — il bias nell'apprendimento: gli individui non scelgono modelli a caso, ma preferenzialmente copiano individui con alta prestazione (payoff bias), alta frequenza (conformity bias), o alta posizione sociale (prestige bias). Questi bias modificano la dinamica delle frequenze esattamente come la selezione naturale modifica le frequenze alleliche.

IV.2 Il conformity bias e la bistabilità culturale

Il conformity bias — la tendenza a copiare la variante maggioritaria con probabilità maggiore della sua frequenza — è uno dei meccanismi più studiati nell'evoluzione culturale. Formalmente, se la frequenza della variante i è f_i , il conformity bias modifica la probabilità di scelta a:

$$P(\text{scelta} = i \mid \text{conformity}) = f_i + D \cdot f_i(1 - f_i)(2f_i - 1)$$

dove $D > 0$ è il parametro del conformity bias. Per $D = 0$ si ha la scelta proporzionale alle frequenze (no bias); per $D > 0$ la frequenza maggioritaria è amplificata.

La dinamica di una popolazione con conformity bias ha la struttura di una bistabilità: il sistema ha tre punti fissi — due stabili ($f = 0$ e $f = 1$, dove una variante ha fissato) e uno instabile ($f = 0.5$, la parità perfetta). Il conformity bias accelera la fissazione rispetto alla sola deriva: il tempo di fissazione per una variante inizialmente frequente è ridotto di un fattore proporzionale a D .

Questa bistabilità ha conseguenze importanti per la coesistenza di culture diverse in popolazioni che si mescolano: il conformity bias produce la soppressione delle varianti minoritarie e l'amplificazione delle maggioritarie — un meccanismo che porta all'omogenizzazione culturale in popolazioni mescolate e alla stabilità delle differenze culturali tra popolazioni isolate. La transizione demografica — la riduzione delle dimensioni delle famiglie nelle società industrializzate — ha ridotto il conformity bias (famiglie con meno figli, esposizione a modelli culturali più diversi) e ha facilitato la diversificazione culturale osservata nelle società moderne.

IV.3 Il modello di Rogers e il paradosso dell'apprendimento sociale

Alan Rogers (1988) dimostrò un paradosso sorprendente nell'evoluzione dell'apprendimento sociale: in una popolazione di individui che usano soltanto l'apprendimento sociale, la fitness media non è superiore a quella di una popolazione che usa soltanto

l'apprendimento individuale. L'apprendimento sociale non aumenta la fitness media della popolazione — soltanto la ridistribuisce tra individui.

La ragione è la seguente. Se tutti gli individui usano l'apprendimento sociale, copiando le strategie comportamentali degli altri, la frequenza delle strategie è determinata dal campionamento casuale delle generazioni precedenti — non dall'accuratezza della risposta all'ambiente corrente. In ambienti variabili, la strategia copiata può essere obsoleta — appropriata per l'ambiente della generazione precedente, non per quello corrente. L'apprendimento individuale, al contrario, produce sempre informazione fresca sull'ambiente attuale.

Il paradosso di Rogers si formalizza come segue. Sia p la frequenza di individui che usano l'apprendimento sociale (SS, social learners) e $1-p$ la frequenza di individui che usano l'apprendimento individuale (IL, individual learners). L'idoneità di ciascun tipo dipende dalla probabilità che la strategia comportamentale sia adattata all'ambiente corrente:

$W_{IL} = 1 - c_{IL}$ (idoneità dell'IL che paga il costo di apprendimento individuale c_{IL}) $W_{SS} = (1 - u) \cdot W_{corretta} + u \cdot W_{obsoleta}$ (idoneità dell'SS: con probabilità $1-u$ l'ambiente non è cambiato e la strategia copiata è corretta, con probabilità u l'ambiente è cambiato e la strategia è obsoleta)

dove u è il tasso di cambiamento ambientale. All'equilibrio evolutivo, $W_{IL} = W_{SS}$ — i due tipi hanno uguale idoneità media. Sostituendo, si ottiene la condizione di equilibrio:

$$p^* = 1 - c_{IL} / (u \cdot \Delta W)$$

dove $\Delta W = W_{corretta} - W_{obsoleta}$ è il vantaggio di avere la strategia corretta. La fitness media della popolazione all'equilibrio è identica a quella di una popolazione pura di IL — l'apprendimento sociale non ha aumentato la fitness media.

Il paradosso di Rogers è risolto da Enquist et al. (2007) introducendo l'apprendimento sociale critico o contingente: individui che usano l'apprendimento sociale soltanto quando la qualità dell'informazione copiata supera una soglia di confidenza, e usano l'apprendimento individuale altrimenti. Questa strategia mista rompe il paradosso e produce fitness media superiore alle strategie pure.

V. Il linguaggio come sistema di trasmissione e le sue distorsioni

V.1 Il telephone game e la degradazione dell'informazione

Il "telefono senza fili" — il gioco in cui una storia o un messaggio viene trasmesso oralmente attraverso una catena di individui — è il modello sperimentale classico della degradazione dell'informazione nella trasmissione culturale. Bartlett (1932), nel suo studio pionieristico "Remembering", mostrò che i soggetti che trasmettono storie di culture diverse tendono a "normalizzare" le storie — sostituire elementi incongruenti con il proprio sistema culturale con elementi familiari, semplificare le strutture narrative complesse, aggiungere elementi di coerenza mancanti nell'originale.

La struttura matematica della degradazione di Bartlett è quella di un attrattore nella ritenzione mnemonica. Sia $s \in S$ la storia originale, $\tilde{s}_i$ la versione ricordata dall'individuo i dopo aver sentito la versione dell'individuo $i-1$. Il processo di trasmissione è:

$$\tilde{s}_i = T(\tilde{s}_{i-1}, c_i, v_i)$$

dove T è la funzione di trasformazione mnemonica, c_i è lo schema culturale dell'individuo i (il suo modello del mondo che guida la codifica e il richiamo), e v_i è il rumore individuale di codifica/decodifica.

In prima approssimazione lineare:

$$\tilde{s}_i = A \cdot \tilde{s}_{i-1} + b + \epsilon_i$$

dove A è la matrice di trasformazione degli schemi culturali (che tira la storia verso le strutture narrative prototipiche della cultura) e b è il bias sistematico verso il prototipo culturale. Il punto fisso di questa dinamica è:

$$s^* = (I - A)^{-1} b$$

che è il prototipo culturale — la storia "canonicamente normale" della cultura i . Dopo un numero sufficiente di trasmissioni, la storia converge al prototipo culturale, indipendentemente dalla storia originale — la storia originale è completamente sostituita dal prototipo. La velocità di convergenza è determinata dall'autovalore dominante di A .

Questo risultato ha un'implicazione fondamentale per la trasmissione delle tradizioni culturali complesse: in assenza di meccanismi di correzione (testi scritti, insegnamento esplicito, rituali di ripetizione precisa), le tradizioni orali convergono progressivamente ai prototipi culturali correnti, perdendo gli elementi idiosincratici e insoliti

dell'originale. La tradizione orale non è un archivio — è un processo di normalizzazione progressiva.

V.2 La fedeltà della scrittura e i suoi limiti

La scrittura è il principale meccanismo culturale per aumentare la fedeltà della trasmissione rispetto alla trasmissione orale. Un testo scritto può essere copiato con fedeltà arbitrariamente alta (nel limite della copia meccanica), permettendo la trasmissione di informazione complessa attraverso generazioni senza il degrado progressivo della trasmissione orale.

Ma anche la scrittura ha limiti di fedeltà strutturali che non dipendono dalla qualità della copia fisica. Il primo limite è l'interpretazione: un testo scritto non contiene tutta l'informazione necessaria per la sua corretta interpretazione — richiede un "background knowledge" (conoscenza di sfondo) che non è esplicitata nel testo e che deve essere trasmessa separatamente. Testi tecnici complessi (istruzioni per costruire uno strumento musicale, ricette culinarie per tecniche avanzate, procedure mediche) richiedono una tradizione pratica viva per la loro corretta interpretazione — senza la pratica, il testo diventa incomprensibile o viene interpretato in modo distorto.

Il secondo limite è il vocabolario: le parole cambiano significato nel tempo, e un testo scritto millecinquecento anni fa usa parole che hanno significati diversi nel presente. La comprensione del testo richiede la ricostruzione del significato originale attraverso l'analisi storica — un processo che introduce inevitabilmente incertezze e distorsioni.

Il terzo limite è la selezione dei testi da copiare: non tutti i testi vengono copiati con uguale probabilità. I testi che corrispondono ai valori e agli interessi del copista (o dei committenti della copia) vengono preferenzialmente copiati, mentre i testi che contraddicono quei valori o che appaiono irrilevanti vengono lasciati deteriorare. Questo produce una selezione sistematica della tradizione scritta — equivalente al conformity bias nella trasmissione orale — che introduce distorsioni cumulative nel corpus trasmesso.

VI. La cognizione distribuita e i limiti della mente estesa

VI.1 Il cervello come parte di un sistema cognitivo distribuito

Edwin Hutchins (1995) ha mostrato attraverso l'analisi della navigazione su navi militari che la cognizione non è confinata nei cervelli degli individui ma è distribuita tra individui, artefatti e ambienti. Il sistema cognitivo rilevante per la navigazione include i marinai, gli strumenti nautici, le mappe, i log di rotta, i protocolli di comunicazione — una rete socio-tecnica che nel suo insieme esibisce proprietà cognitive (pianificazione, errore-correzione, memoria) che nessun componente individuale possiede.

La cognizione distribuita ha due implicazioni per la trasmissione culturale. Primo, la "conoscenza" che viene trasmessa non è soltanto nel cervello degli individui ma è parzialmente esternalizzata in artefatti e strutture ambientali — il know-how per usare una bussola è distribuito tra il cervello dell'operatore (che sa come leggere la bussola) e la bussola stessa (che codifica la relazione tra nord magnetico e lancetta). La perdita di uno dei componenti del sistema cognitivo distribuito — la bussola, o il personale addestrato a usarla — può rendere la conoscenza inaccessibile.

Secondo, la trasmissione culturale è la trasmissione di sistemi cognitivi distribuiti, non soltanto di rappresentazioni individuali. La trasmissione efficace richiede la riproduzione dell'intero sistema — cervelli, artefatti, strutture ambientali, protocolli di interazione — non soltanto dei componenti neurologici. Questo rende la trasmissione culturale molto più costosa e fragile di quanto il modello individuale suggerirebbe.

VI.2 La mente estesa e il problema della continuità

Clark e Chalmers (1998) hanno proposto la tesi della "mente estesa": gli stati cognitivi non sono confinati nei cervelli degli individui ma si estendono agli artefatti esterni quando questi soddisfano tre condizioni funzionali — sono costantemente disponibili, affidabili, e automaticamente consultati (il "parity principle"). Il notebook di Otto, che compensa la sua disfunzione mnemonica, è parte del suo sistema cognitivo nel senso funzionalmente rilevante.

Per la trasmissione culturale, la tesi della mente estesa ha un'implicazione problematica: la trasmissione dell'intelligenza individuale richiede la trasmissione del suo sistema cognitivo esteso — il notebook, il laboratorio, la biblioteca, il network sociale — non soltanto delle capacità neurali. La deprivazione di questi sistemi cognitivi estesi — come avviene nelle catastrofi demografiche e nelle distruzioni deliberate delle biblioteche e dei laboratori — non è una perdita di oggetti materiali ma una perdita di componenti del sistema cognitivo stesso, con effetti sulla competenza cognitiva comparabili a una lesione cerebrale.

La formalizzazione di questo argomento usa la nozione di "complementarità cognitiva". Sia Φ_{int} la capacità cognitiva interna (neurale) di un individuo e Φ_{ext} la capacità cognitiva esterna (artefatti, strutture ambientali). La prestazione cognitiva totale è:

$$\Phi_{\text{tot}} = f(\Phi_{\text{int}}, \Phi_{\text{ext}})$$

dove f non è additiva: le capacità interna ed esterna sono complementari nel senso economico — il rendimento marginale di Φ_{int} è aumentato dalla disponibilità di Φ_{ext} e viceversa. La distruzione di Φ_{ext} riduce la prestazione cognitiva totale più di quanto farebbe la perdita di una risorsa indipendente di uguale valore.

Questo è il meccanismo attraverso cui le catastrofi demografiche che distruggono le strutture sociali e materiali (guerre, epidemie, carestie) producono collassi della complessità tecnologica più profondi di quanto la sola perdita demografica spiegherebbe: si perdono non soltanto gli individui, ma i sistemi cognitivi estesi in cui le competenze erano parzialmente codificate.

VII. La selezione culturale e le sue distorsioni

VII.1 Il bias di contenuto e l'attrattività delle idee

Non tutte le varianti culturali si trasmettono con uguale probabilità: alcune idee, credenze, pratiche e storie sono più "attrattive" — più facilmente apprendibili, più facilmente ricordate, più facilmente condivise — di altre. Dan Sperber (1996) ha chiamato questo fenomeno "epidemiologia delle rappresentazioni": le rappresentazioni culturali si diffondono nelle popolazioni umane attraverso processi analoghi alla trasmissione delle malattie infettive, e la "contagiosità" di una rappresentazione dipende dalle sue proprietà cognitive.

Boyer (2001) ha studiato sistematicamente le proprietà cognitive che rendono le credenze religiose "attrattive" — facilmente apprendibili, ricordate e trasmesse. Il risultato è la "teoria delle violazioni minimali dell'intuizione" (Minimal Counterintuitive Concepts, MCI): le credenze religiose sono credenze che violano minimalmente le aspettative intuitive della fisica ingenua, della biologia ingenua o della psicologia ingenua — agenti con capacità fisiche sovrannaturali (spiriti che attraversano le pareti), oggetti con proprietà biologiche inusuali (alberi che parlano), entità con abilità psicologiche straordinarie (esseri che conoscono tutti i pensieri). Questa struttura — violare una sola

aspettativa intuitiva mentre si soddisfano tutte le altre — produce memorie eccezionalmente durature rispetto alle credenze completamente intuitive (troppo ordinarie per essere memorabili) o completamente controintuitive (troppo strane per essere processate e ricordate).

La formalizzazione è quella dell'ottimizzazione della sorpresa: una rappresentazione è attrattiva quando il suo contenuto viola le aspettative (produce aggiornamento delle credenze, quindi alto impatto mnemonico) senza essere così inattesa da non poter essere processata (non superare la capacità di comprensione del destinatario). In termini di teoria dell'informazione, la rappresentazione ottimalmente attrattiva ha sorpresa

$$S = -\log P(\text{rappresentazione} \mid \text{modello del mondo})$$

che è alta (rappresentazione inattesa) ma non massima (la rappresentazione non supera la capacità di processare informazione inattesa).

Questa struttura produce una selezione delle idee culturali in favore delle rappresentazioni MCI — una selezione che non necessariamente favorisce le rappresentazioni più accurate o più utili, ma quelle cognitivamente più "attrattive". Questo è il meccanismo attraverso cui le credenze religiose e superstizioni si diffondono con tale persistenza nelle culture umane: sono selezionate da pressioni cognitive indipendenti dalla loro veridicità o utilità.

VII.2 Il problema del knowledge-loss e l'erosione della competenza tacita

Polanyi (1958) ha distinto tra conoscenza esplicita (proposizioni, procedure, regole) e conoscenza tacita (il "saper fare" che non si può esprimere in parole). La conoscenza tacita — come si prova il pane per sapere se è cotto, come si ascolta il motore per sentire se è in ordine, come si legge il viso di un interlocutore — è trasmissibile soltanto attraverso l'apprendistato diretto: un maestro che mostra, un allievo che pratica, una relazione che si protrae per mesi o anni.

La perdita della conoscenza tacita è strutturalmente diversa dalla perdita della conoscenza esplicita: la conoscenza esplicita può essere (in principio) ricostruita da testi sopravvissuti; la conoscenza tacita non ha rappresentazione esplicita e scompare con gli individui che la possiedono. La fragilità della trasmissione della conoscenza tacita è massima nelle culture senza scrittura (dove tutta la conoscenza è tacita) e nelle competenze tecniche altamente specializzate (dove la componente tacita è irriducibile).

Formalmente, sia $\Theta = \Theta_{\text{esplicita}} \cup \Theta_{\text{tacita}}$ la conoscenza totale di una cultura, con $\Theta_{\text{esplicita}} \subset \Theta$ la componente esplicitabile e $\Theta_{\text{tacita}} = \Theta \setminus \Theta_{\text{esplicita}}$ la componente tacita. La sopravvivenza di $\Theta_{\text{esplicita}}$ richiede soltanto la sopravvivenza dei documenti fisici (con probabilità p_{doc} per generazione). La sopravvivenza di Θ_{tacita} richiede la continuità della catena di apprendistato (con probabilità p_{appr}^n per una catena di n generazioni). La probabilità di sopravvivenza di Θ nel corso di T generazioni è:

$$P(\Theta_{\text{esplicita}} \text{ sopravvive per } T \text{ generazioni}) = p_{\text{doc}}^T$$

$$P(\Theta_{\text{tacita}} \text{ sopravvive per } T \text{ generazioni}) = p_{\text{appr}}^{(nT)}$$

con $p_{\text{appr}} < 1$ (ogni trasmissione ha probabilità positiva di interruzione). Per $T \rightarrow \infty$, $P(\Theta_{\text{tacita}} \text{ sopravvive}) \rightarrow 0$ — la conoscenza tacita è inevitabilmente perduta nel lungo periodo, qualunque sia $p_{\text{appr}} > 0$. Solo $p_{\text{appr}} = 1$ (trasmissione perfetta senza mai interruzione) garantisce la sopravvivenza infinita — una condizione non realizzabile.

VIII. La dinamica evolutiva dei sistemi di credenze

VIII.1 La selezione delle idee in popolazioni strutturate

L'evoluzione culturale non opera su individui isolati ma su individui immersi in strutture sociali — famiglie, clan, villaggi, nazioni — che determinano chi interagisce con chi e quindi chi apprende da chi. La struttura della rete sociale determina la velocità di diffusione delle innovazioni culturali e la possibilità di coesistenza di varianti culturali diverse.

Per una rete sociale strutturata in cui ciascun individuo ha k vicini (da cui può apprendere), la condizione per la diffusione di una variante culturale vantaggiosa (con vantaggio selettivo s rispetto alla variante esistente) è analoga alla condizione di Ohtsuki per la cooperazione su grafi (discussa nella Monografia XIII del ciclo precedente):

$$s > 1/k$$

La variante si diffonde se il suo vantaggio selettivo supera l'inverso del numero di vicini. In reti con bassa connettività (k piccolo, come nelle comunità rurali isolate), soltanto varianti con vantaggio molto alto possono diffondersi. In reti con alta connettività (k grande, come nelle

metropoli moderne con accesso a internet), anche varianti con vantaggio piccolo possono diffondersi.

Questa struttura ha un'implicazione per la diversità culturale: le reti a bassa connettività mantengono la diversità culturale attraverso l'isolamento (le varianti si fissano localmente), mentre le reti ad alta connettività convergono all'omogenizzazione culturale (le varianti più attrattive o vantaggiose si diffondono globalmente). La globalizzazione culturale — la convergenza verso culture sempre più simili nelle società ad alta connettività — è la conseguenza diretta di questo meccanismo.

IX. Conclusioni: il cricchetto e le sue crepe

Il tema unificante di questa monografia è la fragilità strutturale della trasmissione culturale — la tendenza inevitabile alla degradazione della complessità e della fedeltà informativa, che il modello ingenuo del "cricchetto" di Tomasello non cattura completamente.

Il cricchetto ruota, ma si inceppa: la fedeltà della trasmissione decresce con la distanza temporale dalla sorgente (trasmissione orale), con la riduzione demografica della rete di apprendimento (modello di Henrich), con la perdita degli artefatti cognitivi che esternalizzano parte della conoscenza (cognizione distribuita), con la selezione delle idee in favore dell'attrattività cognitiva piuttosto che dell'accuratezza (epidemiologia delle rappresentazioni). Ogni generazione trasmette qualcosa di ciò che ha ricevuto, ma meno di ciò che ha ricevuto — a meno che non intervenga l'innovazione attiva o la trasmissione di alta fedeltà (scrittura, insegnamento formale, apprendistato prolungato) a compensare la perdita per deriva.

Il paradosso di Rogers dimostra che l'apprendimento sociale puro non aumenta la fitness media — richiede una combinazione ottimale con l'apprendimento individuale per produrre i suoi benefici. Il modello di Henrich dimostra che la complessità tecnologica richiede una soglia demografica minima — non si può costruire una civiltà con una rete sociale troppo piccola. Il modello di Bartlett dimostra che la trasmissione orale è attrattore verso i prototipi culturali — le storie si normalizzano, le tecniche si semplificano, la diversità si riduce.

Questi risultati sono la *pars destruens* biologica della trasmissione culturale: le impossibilità strutturali che limitano il potere del cricchetto. La *pars costruens* — come i sistemi culturali umani hanno trovato soluzioni parziali a queste impossibilità attraverso la scrittura, le

istituzioni, le scuole, la specializzazione cognitiva distribuita — è il tema che le monografie successive affrontano dal lato della costruzione.

Monografia VIII

Il principio di energia libera e la cognizione predittiva

La mente come motore inferenziale, la minimizzazione della sorpresa e il fondamento unificante della neurobiologia

I. Il problema di Helmholtz e la sua soluzione frattale

Hermann von Helmholtz propose nel 1867 una visione della percezione che per un secolo rimase più metafora filosofica che teoria scientifica rigorosa: la percezione è un'inferenza inconscia (*unbewusste Schlüsse*). Il cervello non registra passivamente gli stimoli sensoriali; li interpreta attivamente, costruendo la migliore ipotesi sulle cause del pattern sensoriale osservato. La percezione è la conclusione di un sillogismo in cui le premesse sono i dati sensoriali e i principi generali dell'organizzazione del mondo (i *prior bayesiani*), e la conclusione è la rappresentazione percettiva.

Questa intuizione ha trovato la sua formalizzazione scientifica precisa nel principio di energia libera di Karl Friston (2005, 2010, 2019) — una teoria dell'ampiezza e della profondità straordinarie che propone un principio variazionale unico come fondamento di tutta la neurobiologia cognitiva: i sistemi cognitivi biologici minimizzano la propria energia libera variazionale, e questa minimizzazione è sufficiente a spiegare la percezione, l'azione, l'attenzione, l'apprendimento, la memoria, le emozioni, la cognizione sociale, lo sviluppo e la patologia.

L'ambizione del programma di Friston è comparabile, nelle scienze cognitive, all'ambizione del principio di minima azione nella fisica: trovare il principio variazionale unico da cui tutte le leggi derivano come conseguenze. La differenza è che il principio di minima azione è dimostrato matematicamente corrispondere alle equazioni di Hamilton-Jacobi; il principio di energia libera è un framework teorico

ancora in sviluppo, con applicazioni ben confermate in certi domini e speculazioni in altri. Questa monografia sviluppa il framework con la massima rigorosità matematica che il suo stato attuale permette, distinguendo le parti dimostrate dalle congetture.

II. L'energia libera variazionale: derivazione formale

II.1 Il problema dell'inferenza bayesiana approssimata

Il punto di partenza del principio di energia libera è un problema fondamentale dell'inferenza bayesiana: il calcolo della distribuzione posteriore $p(\theta | x)$ è in generale computazionalmente intrattabile perché richiede il calcolo dell'evidenza (evidence o marginal likelihood):

$$p(x) = \int p(x | \theta) p(\theta) d\theta$$

Questo integrale è trattabile analiticamente soltanto per coppie specifiche di verosimiglianza e prior (le famiglie coniugate, come Gaussiana-Gaussiana, Binomiale-Beta, Poisson-Gamma), e intrattabile per tutte le altre. Per reti neurali biologiche con milioni di parametri, il calcolo diretto di $p(x)$ è computazionalmente impossibile.

L'inferenza variazionale — sviluppata in statistica da Jordan et al. (1999) e in neuroscienze da Friston (2005) — affronta questo problema approssimando la distribuzione posteriore vera $p(\theta | x)$ con una distribuzione approssimante $q(\theta)$ tratta da una famiglia parametrica trattabile (ad esempio distribuzioni gaussiane fattorizzabili). Il problema di inferenza diventa il problema di ottimizzazione:

$$q^*(\theta) = \operatorname{argmin}_{\{q \in Q\}} D_{\text{KL}}(q(\theta) || p(\theta | x))$$

dove D_{KL} è la divergenza di Kullback-Leibler:

$$D_{\text{KL}}(q(\theta) || p(\theta | x)) = \int q(\theta) \log [q(\theta) / p(\theta | x)] d\theta$$

La divergenza KL è sempre non-negativa e si annulla se e soltanto se $q(\theta) = p(\theta | x)$. Minimizzare D_{KL} è equivalente a trovare la migliore approssimazione di $p(\theta | x)$ nella famiglia Q .

II.2 La decomposizione dell'energia libera variazionale

Attraverso una semplice manipolazione algebrica, la divergenza KL si decompone in:

$$D_{KL}(q \parallel p(\theta \mid x)) = -ELBO + \log p(x)$$

dove ELBO (Evidence Lower BOund, o "energia libera variazionale" in notazione di Friston) è:

$$F = -ELBO = E_q[\log q(\theta)] - E_q[\log p(x, \theta)] = D_{KL}(q(\theta) \parallel p(\theta)) - E_q[\log p(x \mid \theta)]$$

La prima riga dà la definizione operativa: F è l'opposto dell'ELBO. La seconda riga decompone F in due termini:

$$F = D_{KL}(q(\theta) \parallel p(\theta)) - E_q[\log p(x \mid \theta)]$$

Il primo termine $D_{KL}(q(\theta) \parallel p(\theta))$ è la divergenza tra la distribuzione approssimante q e il prior $p(\theta)$ — misura quanto q si discosta dal prior, ovvero quanta informazione aggiuntiva sul prior il modello "porta". Il secondo termine $-E_q[\log p(x \mid \theta)]$ è l'energia attesa — misura quanto bene il modello predice i dati x sotto la distribuzione q .

Poiché $D_{KL}(q \parallel p(\theta \mid x)) \geq 0$, si ha:

$$F \geq -\log p(x) = -\log p(x)$$

L'energia libera variazionale è un upper bound sull'energia libera vera $-\log p(x) = -\log(\text{evidenza})$. Minimizzare F è equivalente a massimizzare un lower bound sull'evidenza (l'ELBO) — da qui il nome Evidence Lower BOund.

Crucialmente:

$$F = -\log p(x) + D_{KL}(q \parallel p(\theta \mid x))$$

La differenza tra F e $-\log p(x)$ è esattamente $D_{KL}(q \parallel p(\theta \mid x))$, che misura quanto la distribuzione approssimante q si discosta dalla vera distribuzione posteriore. Minimizzare F rispetto a q riduce simultaneamente $D_{KL}(q \parallel p(\theta \mid x))$ (migliora l'approssimazione della posteriore) e, quando il modello generativo $p(x, \theta)$ ha parametri liberi, massimizza $\log p(x)$ (aumenta l'evidenza del modello — il suo fit ai dati).

II.3 La riscrittura in termini di sorpresa e accuratezza

Friston propone una riscrittura dell'energia libera che rivela la sua struttura neuroscientifica. Sviluppando il termine di log-verosimiglianza:

$$E_q[\log p(x \mid \theta)] = -E_q[-\log p(x \mid \theta)] = -(\text{Sorpresa attesa})$$

Il termine $-\log p(x | \theta)$ è la sorpresa del dato x sotto il modello con parametri θ — quanto il modello è sorpreso dall'osservazione x . L'accuratezza è l'opposto della sorpresa media:

$$\text{Accuratezza} = E_q[\log p(x | \theta)] = -\text{Sorpresa attesa}$$

Quindi:

$$F = \text{Complessità} - \text{Accuratezza}$$

dove la Complessità è $D_{\text{KL}}(q || p(\theta))$ e l'Accuratezza è $E_q[\log p(x | \theta)]$.

Un buon modello generativo ha alta accuratezza (predice bene i dati) e bassa complessità (non si discosta troppo dal prior — evita l'overfitting). La minimizzazione di F bilancia automaticamente questi due obiettivi — è una formalizzazione del principio di parsimonia (rasoio di Occam bayesiano): tra i modelli che predicono ugualmente bene i dati, preferire il più semplice.

III. La percezione come inferenza variazionale

III.1 Il modello generativo del mondo come struttura gerarchica

Nel framework di Friston, il cervello implementa un modello generativo del mondo — una distribuzione di probabilità congiunta $p(x, \theta, \varphi, \dots)$ su dati sensoriali x , cause dirette θ , cause di livello superiore φ , e così via per quante gerarchie il sistema include. Questo modello generativo è gerarchico: le variabili di livello superiore producono le variabili di livello inferiore attraverso relazioni causali che il modello specifica.

La struttura gerarchica del modello generativo è:

$$\varphi \sim p(\varphi | \varphi_{+1}) \text{ (livello } n \text{ causato dal livello } n+1) \dots \varphi_1 \sim p(\varphi_1 | \varphi_2) \text{ (livello 1: cause prossime)} \\ x \sim p(x | \varphi_1) \text{ (dati sensoriali causati dal livello 1)}$$

Il modello generativo completo è:

$$p(x, \varphi_1, \dots, \varphi) = p(x | \varphi_1) \cdot \prod_i p(\varphi_i | \varphi_{i+1})$$

L'inferenza percettiva è il calcolo della distribuzione posteriore sulle cause $\varphi_1, \dots, \varphi$ dato il dato sensoriale x . Poiché questo calcolo è intrattabile esattamente, il cervello implementa l'inferenza variazionale: mantiene una distribuzione approssimante $q(\varphi_1, \dots, \varphi)$ che minimizza F .

La distribuzione approssimante q , per ragioni di trattabilità, assume la fattorizzazione di campo medio:

$$q(\varphi_1, \dots, \varphi) = \prod_i q(\varphi_i)$$

(le variabili di livelli diversi sono approssimativamente indipendenti in q). Questa è la "mean field approximation" dell'inferenza variazionale.

III.2 Il predictive coding come implementazione neurale

La minimizzazione di F attraverso l'aggiornamento di q dà le equazioni del "predictive coding" — il meccanismo attraverso cui le aree corticali gerarchiche comunicano. L'aggiornamento ottimale per $q(\varphi_i)$ nella fattorizzazione di campo medio è:

$$\partial F / \partial \mu_i = \varepsilon_i - \varepsilon_{i+1}$$

dove μ_i è la media di $q(\varphi_i)$, $\varepsilon_i = \mu_i - \hat{g}_i(\mu_{i+1})$ è l'errore di previsione al livello i (la differenza tra il valore corrente μ_i e la previsione $\hat{g}_i(\mu_{i+1})$ generata dal livello superiore $i+1$), e ε_{i+1} è l'errore di previsione che il livello i causa al livello inferiore $i-1$.

Il punto fisso di questa equazione — quando $\partial F / \partial \mu_i = 0$ — dà il predictive coding:

$$\varepsilon_i = \varepsilon_{i+1}$$

Gli errori di previsione si propagano su su nella gerarchia, e le previsioni si propagano verso il basso. L'implementazione neurale proposta da Rao e Ballard (1999) e Friston (2005) è:

Neuroni superficiali delle aree corticali: trasmettono gli errori di previsione verso l'alto (segnali bottom-up, dalle aree sensoriali verso le aree associative)

Neuroni profondi delle aree corticali: trasmettono le previsioni verso il basso (segnali top-down, dalle aree associative verso le aree sensoriali)

I neuroni superficiali e profondi sono anatomicamente distinti nelle lamine corticali (strati I-III vs. strati V-VI), con connessioni bottom-up prevalentemente attraverso i layer granulari (strato IV) e connessioni

top-down attraverso i layer extragranulari — una struttura anatomica ben documentata che corrisponde alla predizione del predictive coding.

La percezione è il processo di minimizzazione degli errori di previsione attraverso l'aggiornamento delle previsioni di livello superiore: le rappresentazioni ai livelli superiori si aggiornano per ridurre gli errori di previsione nei livelli inferiori. La percezione stabile — quando si percepisce chiaramente un oggetto — corrisponde a uno stato in cui gli errori di previsione sono minimizzati a tutti i livelli della gerarchia.

IV. L'azione come inferenza attiva

IV.1 La minimizzazione dell'energia libera attraverso l'azione

Il principio di energia libera non si applica soltanto alla percezione (aggiornamento delle credenze per ridurre F) ma anche all'azione. Friston (2010) propone che l'azione sia anch'essa un processo di minimizzazione di F , ma attraverso un meccanismo diverso: invece di aggiornare le credenze per renderle compatibili con i dati sensoriali, l'agente cambia i dati sensoriali per renderli compatibili con le proprie credenze.

Formalmente, se l'agente ha un'aspettativa (una credenza forte) su come dovrebbe essere il mondo — una distribuzione prior q_{target} — e il mondo attuale produce dati sensoriali x che differiscono da quelli attesi, l'agente può minimizzare F sia aggiornando q_{target} (adattare le credenze al mondo) sia cambiando x attraverso l'azione (adattare il mondo alle credenze).

La scelta tra percettuale (aggiornare le credenze) e azione (cambiare il mondo) dipende dalla struttura di F : se la varianza del prior q_{target} è bassa (alta certezza) e la varianza del modello sensoriale è alta (segnale rumoroso), aggiornare le credenze è inefficiente e l'azione è preferita.

Questa asimmetria formalizza l'intuizione gibsoniana dell'inferenza attiva: gli agenti non reagiscono passivamente all'ambiente ma lo modificano attivamente per renderlo conforme alle proprie aspettative. Il gatto che si avvicina alla preda non sta "costruendo una rappresentazione accurata della preda" — sta eseguendo azioni che riducono la sorpresa rispetto al proprio modello predittivo "ho la preda".

IV.2 L'azione come minimizzazione degli errori propriocettivi

L'implementazione neurale dell'azione come inferenza attiva si basa sulla seguente osservazione: i motoneuroni non ricevono soltanto segnali dalle aree motorie corticali, ma anche segnali propriocettivi dal midollo spinale. Nel modello di predictive coding, i motoneuroni implementano le equazioni del predictive coding per il segnale propriocettivo: minimizzano l'errore tra la posizione attesa dell'arto (generata dal modello predittivo corticale) e la posizione attuale (misurata dai fusi neuromuscolari).

L'equazione di aggiornamento per l'azione a è:

$$da/dt = -\partial F/\partial a = \partial F/\partial x \cdot \partial x/\partial a$$

dove $\partial F/\partial x$ è il gradiente dell'energia libera rispetto ai dati sensoriali (legato agli errori di previsione) e $\partial x/\partial a$ è il gradiente dell'effetto dell'azione sui dati sensoriali (il modello della dinamica motoria). L'azione si aggiorna nella direzione che riduce F — ossia nella direzione che riduce gli errori di previsione propriocettivi.

Questo framework unifica il controllo motorio e la percezione nel medesimo principio: entrambi minimizzano F , l'uno aggiornando le rappresentazioni interne, l'altro cambiando i dati sensoriali. La struttura è quella di un controllo a feedback ottimale con un elemento predittivo: il sistema non reagisce agli errori attuali, ma agli errori predetti dai livelli superiori della gerarchia.

V. L'attenzione come precisione

V.1 La precisione come modulazione degli errori di previsione

Nel framework del predictive coding, non tutti gli errori di previsione hanno lo stesso peso nell'aggiornamento delle credenze: gli errori di previsione di modalità sensoriali più affidabili (in certi contesti) devono essere pesati più di quelli di modalità meno affidabili. Questo peso — la precisione dell'errore di previsione — è il correlato computazionale dell'attenzione.

Formalmente, nel modello gerarchico con errori di previsione ε_i al livello i , la minimizzazione di F introduce un termine di precisione:

$$F = \sum_i \Pi_i \cdot \|\epsilon_i\|^2 + \log \det(\Sigma_i)$$

dove $\Pi_i = \Sigma_i^{-1}$ è la matrice di precisione (inversa della covarianza) degli errori al livello i e Σ_i è la covarianza degli errori al livello i . La precisione Π_i modula l'influenza degli errori di previsione del livello i sull'aggiornamento delle credenze: errori di alta precisione producono aggiornamenti forti; errori di bassa precisione producono aggiornamenti deboli.

L'attenzione è, in questa formalizzazione, la modulazione della precisione degli errori di previsione: dirigere l'attenzione verso una modalità sensoriale o verso una parte del campo visivo equivale ad aumentare la precisione degli errori di previsione per quella modalità o quella regione. La conseguenza neuroscientifica è che l'attenzione dovrebbe essere implementata attraverso il guadagno dei neuroni superficiali (che trasmettono gli errori di previsione) — una previsione confermata da studi fisiologici che mostrano che l'attenzione aumenta il guadagno neurale nelle aree visive sensoriali.

Il neuromodulatore acetilcolina (ACh) è il candidato molecolare principale per la codifica della precisione attenzionale. L'ACh è rilasciata dal nucleo basale di Meynert e dall'area setale verso tutta la corteccia, e la sua concentrazione è correlata con il livello di attenzione. Farmacologicamente, gli antagonisti muscarinici (che bloccano la segnalazione ACh) riducono la capacità attenzionale; gli agonisti la aumentano. Nel modello di precisione, l'ACh modula il guadagno dei neuroni superficiali — regola la "credibilità" degli errori di previsione.

V.2 I neuromodulatori come segnali di meta-precisione

Il framework di Friston (2012) propone che i diversi neuromodulatori codifichino diversi aspetti della precisione in diversi contesti:

- Dopamina: codifica la precisione delle previsioni di ricompensa (errore di predizione della ricompensa) — la certezza che un'azione produce una ricompensa attesa.
- Acetilcolina: codifica la precisione degli errori di previsione sensoriali — l'affidabilità dei segnali sensoriali nell'ambiente corrente.
- Noradrenalina: codifica la volatilità ambientale — la velocità con cui l'ambiente cambia, che determina quanto rapidamente le credenze devono essere aggiornate.
- Serotonina: codifica l'orizzonte temporale delle conseguenze delle azioni — il peso relativo delle conseguenze immediate rispetto a quelle ritardate.

Questa mappatura — ciascun neuromodulatore codifica una dimensione diversa della struttura della precisione — è una previsione teorica che trova supporto in molteplici studi farmacologici e clinici. Le malattie mentali, in questo framework, sono disfunzioni della codifica della precisione: la schizofrenia è un'aberrazione della precisione degli errori di previsione sensoriali (percezione di errori di previsione che non riflettono perturbazioni del mondo — le allucinazioni); il disturbo depressivo maggiore è un'eccessiva precisione degli errori di previsione negativi (il pessimismo come prior troppo preciso sulla negatività degli esiti).

VI. L'apprendimento come aggiornamento del modello generativo

VI.1 La struttura gerarchica dell'apprendimento

Nel framework di Friston, l'apprendimento è l'aggiornamento lento dei parametri del modello generativo — la struttura del modello che specifica le relazioni causali tra le variabili. A differenza della percezione (aggiornamento veloce delle credenze sui valori delle variabili) e dell'azione (cambiamento dei dati sensoriali), l'apprendimento è l'aggiornamento della struttura del modello stesso.

Formalmente, siano θ i parametri del modello generativo $p(x, y; \theta)$. L'apprendimento è:

$$d\theta/dt = -\eta \cdot \partial F / \partial \theta = \eta \cdot \partial \text{ELBO} / \partial \theta$$

dove η è il tasso di apprendimento e la derivata è il gradiente dell'ELBO (negativo dell'energia libera) rispetto ai parametri. Questo è l'algoritmo di massimizzazione dell'ELBO (EM, o varianti stocastiche come il gradient descent variazionale) che massimizza l'evidenza del modello.

La struttura gerarchica dell'apprendimento corrisponde alla gerarchia temporale: i parametri dei livelli superiori (che codificano strutture più astratte) cambiano più lentamente dei parametri dei livelli inferiori (che codificano dettagli più concreti). Questa struttura di apprendimento a tassi multipli è formalizzata nell'apprendimento gerarchico bayesiano:

θ cambia a velocità η_n con $\eta_1 \gg \eta_2 \gg \dots \gg \eta$

La struttura temporale dell'apprendimento ha un correlato biologico preciso: le diverse aree corticali hanno costanti di tempo diverse nella plasticità sinaptica — l'ippocampo apprende rapidamente (formazione

di ricordi episodici in un solo trial), le aree corticali associative lentamente (consolidazione nel corso di settimane e mesi), le aree sensoriali primarie molto lentamente (cambiamenti strutturali che richiedono periodi critici o deafferentazione prolungata).

VI.2 La codifica predittiva come spiegazione delle illusioni

Le illusioni percettive — il cubo di Necker, la maschera concava, il flash-lag effect — sono spesso citate come dimostrazioni della natura costruttiva della percezione. Nel framework del predictive coding, le illusioni non sono anomalie da spiegare separatamente: sono la predizione diretta del modello che le cause più probabili delle evidenze sensoriali vengono percepite, anche quando l'evidenza sensoriale è ambigua o contraddittoria.

L'illusione della maschera concava (il fenomeno per cui una maschera con il volto all'interno sembra convessa invece che concava) è un esempio paradigmatico. Il modello generativo corticale ha un prior molto forte sulla convessità dei volti (nella nostra esperienza quotidiana, quasi tutti i volti che incontriamo sono convessi — non concavi). Questo prior, quando applicato all'immagine ambigua di una maschera vista dall'interno, produce la percezione di un volto convesso — nonostante i dati sensoriali (l'ombreggiatura, la prospettiva) siano coerenti con la concavità.

Formalmente, il posterior $p(\text{convessa} \mid \text{immagine})$ è dominato dal prior $p(\text{convessa})$ rispetto alla verosimiglianza $p(\text{immagine} \mid \text{concava})$ perché il prior è molto più preciso (molta più evidenza nella vita di quante immagini di maschere concave si siano viste). Il sistema percepisce il volto convesso perché questa è la credenza con la più alta probabilità posteriore — nonostante la maschera sia effettivamente concava.

VII. Il principio di energia libera e i sistemi biologici: generalizzazione

VII.1 L'auto-organizzazione biologica come minimizzazione di F

Friston (2013) ha proposto che il principio di energia libera non si applichi soltanto ai sistemi neurali ma a qualsiasi sistema biologico che mantiene la propria organizzazione contro la tendenza termodinamica verso il disordine (verso l'equilibrio). Un sistema biologico — una

cellula, un organismo, un'ecologia — è un sistema che resiste alla tendenza verso l'equilibrio termodinamico. Come fa?

La proposta di Friston è che qualsiasi sistema con un "confine di Markov" (Markov blanket) — un insieme di stati che schermano il sistema interno dall'ambiente esterno — tende a minimizzare la sorpresa sui propri stati sensoriali (gli stati al confine di Markov). Questo è equivalente a minimizzare F .

Il confine di Markov è un insieme di stati B tale che gli stati interni μ sono condizionalmente indipendenti dall'ambiente η dato B :

$$p(\mu \mid \eta, B) = p(\mu \mid B)$$

Il confine "protegge" l'interno dall'esterno: l'ambiente può influenzare l'interno soltanto attraverso il confine. Per un neurone, il confine di Markov sono la membrana cellulare e le sinapsi; per un organo, i tessuti connettivi; per un organismo, la pelle e gli organi di senso; per un cervello, il cranio e il sistema periferico nervoso.

La tesi di Friston è che qualsiasi sistema che mantiene un confine di Markov nel tempo deve essere strutturato in modo tale da minimizzare la sorpresa sugli stati del confine — perché sistemi che non minimizzano la sorpresa si dissolvono nell'ambiente (perdono il confine di Markov). L'auto-organizzazione biologica è la conseguenza di questa pressione: soltanto i sistemi che minimizzano F sopravvivono come entità distinte dall'ambiente.

VII.2 I limiti della generalizzazione: la critica di Colombo e Seriès

La generalizzazione del principio di energia libera da sistemi neurali a sistemi biologici in generale ha ricevuto critiche formali importanti (Colombo e Seriès, 2012; Linson et al., 2018). Il problema centrale è che la minimizzazione di F è una condizione necessaria ma non sufficiente per l'auto-organizzazione biologica: molti sistemi fisici (cristalli, magneti, gas a bassa temperatura) mantengono confini di Markov e non sono biologici, né si comportano in modo cognitivo.

La critica formale è la seguente. Il principio di energia libera afferma che un sistema con confine di Markov che persiste nel tempo deve minimizzare F sui propri stati. Ma questo è una tautologia: se definiamo F come $-\log p(\text{stati al confine} \mid \text{modello})$, allora un sistema che persiste è per definizione in stati a bassa sorpresa (alta probabilità) — semplicemente perché gli stati ad alta sorpresa corrispondono alla dissoluzione del sistema. La minimizzazione di F non è un principio

esplicativo; è una ridescrittura della persistenza biologica in linguaggio variazionale.

Friston (2019) risponde che la forza esplicativa del principio non è nella tautologia (qualsiasi sistema persistente minimizza F) ma nella struttura che il principio impone: il sistema deve avere un modello generativo del mondo, deve implementare l'inferenza variazionale, deve avere un confine di Markov — strutture che possono essere studiate empiricamente e che producono previsioni testabili. La valutazione critica di questa risposta rimane aperta nella letteratura.

VIII. Le patologie cognitive come disfunzioni della minimizzazione di F

VIII.1 La schizofrenia come aberrazione della precisione

Il framework del predictive coding ha prodotto la teoria computazionale della schizofrenia più influente degli ultimi vent'anni (Fletcher e Frith, 2009; Adams et al., 2013). La tesi centrale è che la schizofrenia è causata da un'aberrazione della codifica della precisione: i segnali di errore di previsione hanno troppa precisione (troppa influenza sull'aggiornamento delle credenze) anche quando non riflettono perturbazioni del mondo.

Formalmente, nella schizofrenia:

$$\Pi_i^* = \Pi_i + \delta\Pi$$

dove $\delta\Pi > 0$ è un eccesso di precisione (mediated from dopaminergic dysregulation: la dopamina aumenta la precisione degli errori di previsione). Questo produce due famiglie di sintomi.

I sintomi positivi (allucinazioni, deliri) emergono quando errori di previsione interni (generati dal modello predittivo) sono trattati come errori di previsione reali (generati da perturbazioni del mondo). Se il sistema attribuisce alta precisione a segnali di errore che non corrispondono a eventi reali — ad esempio le attivazioni casuali nel sistema uditivo durante il silenzio — queste attivazioni vengono interpretate come segnali del mondo (voci, suoni) invece che come rumori interni.

I sintomi negativi (alogia, avolizione, anedonia) emergono dalla diffusione dell'aberrazione della precisione ai segnali di previsione della

ricompensa (dopaminergici): se la precisione delle previsioni di ricompensa è ridotta, le azioni non producono segnali di errore di ricompensa adeguati, l'apprendimento per rinforzo si indebolisce, e la motivazione a perseguire ricompense cade.

La conferma farmacologica è coerente: gli antipsicotici efficaci sono tipicamente antagonisti dei recettori della dopamina (D2R) — riducono la precisione degli errori di previsione mediata dalla dopamina — mentre gli psicostimolanti (anfetamine, cocaina) che aumentano la dopamina producono sintomi simili alla schizofrenia positiva.

VIII.2 La depressione come eccesso di precisione dei prior negativi

La depressione, nel framework del predictive coding, è una disfunzione complementare alla schizofrenia: non un eccesso di precisione degli errori di previsione (schizofrenia) ma un eccesso di precisione dei prior negativi — credenze troppo certe e negative sullo stato del mondo.

Formalmente, nella depressione:

$$q_{\text{depressione}}(\theta) = N(\mu_{\text{neg}}, \sigma^2_{\text{neg}}) \text{ con } \sigma^2_{\text{neg}} \ll \sigma^2_{\text{normale}}$$

Il prior depresso ha varianza molto bassa: la persona crede con alta certezza che il futuro sarà negativo, che le proprie azioni non produrranno esiti positivi, che nulla cambierà. Questa alta precisione del prior negativo produce due effetti: riduce la sensibilità ai segnali positivi (gli errori di previsione positivi sono soppressi dalla precisione del prior negativo) e riduce l'esplorazione comportamentale (la certezza del futuro negativo riduce l'incentivo ad agire).

La struttura del prior depresso nella teoria di Friston corrisponde alla "learned helplessness" di Seligman (1975): la certezza appresa che le proprie azioni non hanno effetti sull'ambiente. Ma mentre Seligman la descrive comportamentalmente, Friston la formalizza come eccesso di precisione del prior negativo — una struttura che permette previsioni sulla risposta ai trattamenti: gli antidepressivi efficaci (SSRI, ketamina) dovrebbero ridurre la precisione dei prior negativi o aumentare la sensibilità agli errori di previsione positivi.

IX. I limiti del framework: ciò che il principio di energia libera non spiega

IX.1 Il problema difficile resta aperto

Per quanto potente e unificante, il principio di energia libera non affronta il problema difficile della coscienza (discusso nella Monografia IV). Friston (2010, 2019) propone che la minimizzazione di F sia ciò che i sistemi biologici fanno, ma questo non spiega perché vi sia qualcosa che è come farlo. Una macchina che minimizza F — un termostato, un sistema di controllo automatico, un robot bayesiano — fa la stessa cosa del cervello (computazionalmente), ma nessuno pensa che provi esperienza soggettiva.

Friston è consapevole di questo: il principio di energia libera è proposto come spiegazione funzionale della mente, non come soluzione al problema difficile. La domanda di perché la minimizzazione di F sia accompagnata da esperienza soggettiva nei sistemi biologici — una domanda che coincide con l'explanatory gap di Chalmers — rimane aperta anche nel framework del predictive coding.

IX.2 La circolarità del modello generativo

Un problema formale del framework è la questione dell'origine del modello generativo: il sistema minimizza F rispetto al proprio modello generativo interno $p(x, \varphi; \theta)$, ma da dove viene questo modello? Il framework presuppone che il modello esista — ed è corretto nel descrivere come viene aggiornato (attraverso la minimizzazione di F) — ma non spiega come la struttura del modello sia stata costruita nella prima parte dello sviluppo.

La risposta naturale è che il modello è costruito a sua volta attraverso la minimizzazione di F nelle fasi precoci dello sviluppo — ma questo presuppone un modello anche nelle fasi precoci, il che sposta il problema all'indietro senza risolverlo. Il bootstrap del modello generativo — come un sistema cognitivo inizia a costruire il proprio modello del mondo da zero, con un modello iniziale che per definizione è vuoto o casuale — è un problema aperto nel framework.

X. Conclusioni: il principio di energia libera come pars costruens biologica

Il principio di energia libera è, nella tassonomia di questo ciclo, la risposta costruttiva biologica ai problemi di impossibilità analizzati nella pars destruens. Dove la Monografia I mostrava i limiti della rappresentazione neurale, il predictive coding mostra come il cervello

risolve il problema della classificazione attraverso il modello generativo gerarchico. Dove la Monografia III mostrava l'impossibilità del frame problem, il principio di energia libera mostra come il sistema cognitivo gestisce l'irrelevanza attraverso la struttura dei prior e della precisione. Dove la Monografia IV mostrava l'explanatory gap, il principio di energia libera mostra fino a dove può arrivare la spiegazione funzionale — e dove si ferma.

La profondità del programma di Friston è che unifica sotto un unico principio variazionale fenomeni che la neuroscienza tradizionale trattava separatamente: percezione, azione, attenzione, apprendimento, sviluppo, patologia. Questo non significa che il programma sia completo o dimostrato — il framework è work in progress, con previsioni non ancora verificate e con critiche formali legittime che ne circoscrivono la portata. Ma come programma di ricerca — come la mappa di un territorio da esplorare — il principio di energia libera offre la struttura più ambiziosa e coerente oggi disponibile per una teoria unificata della cognizione biologica.

Monografia IX

L'inferenza attiva e la causalità incarnata

Il corpo come strumento epistemico, la profezia che si autorealizza e la struttura causale dell'agency biologica

I. Il paradosso del corpo che non sa di essere conosciuto

La tradizione filosofica occidentale ha trattato il corpo come oggetto della conoscenza — qualcosa che il soggetto conosce nello stesso modo in cui conosce una sedia o una montagna, dall'esterno, attraverso l'osservazione. Questa tradizione, che raggiunge il suo apice nel dualismo cartesiano e persiste in forme attenuate nel funzionalismo cognitivista contemporaneo, ha prodotto un paradosso che Merleau-Ponty (1945) identificò con precisione: il corpo è al tempo stesso soggetto e oggetto, il medium di ogni esperienza e allo stesso tempo un'entità nel mondo. Non si può "guardare" il corpo proprio nello stesso modo in cui si guarda un oggetto esterno — perché qualsiasi atto di guardare è già condizionato dall'architettura corporea che lo rende possibile.

Il principio di energia libera della Monografia VIII ha reso questo paradosso formalmente preciso: l'agente non è esterno al sistema causale che osserva ma ne è una parte strutturale — il nodo Alg del grafo causale di Yudkowsky-Soares, ma nella sua versione biologica continua e incarnata. L'inferenza attiva è il meccanismo attraverso cui questa posizione paradossale — essere simultaneamente l'osservatore e l'osservato, l'agente e il paziente — viene gestita dal sistema nervoso biologico: non risolvendo il paradosso in favore di uno dei poli, ma sfruttando la struttura circolare per costruire un sistema di controllo che è al tempo stesso stabile e adattivo.

Questa monografia sviluppa la teoria dell'inferenza attiva come teoria dell'agency biologica incarnata — il meccanismo attraverso cui gli agenti biologici non soltanto elaborano informazione sul mondo ma lo modificano attivamente per renderlo conforme alle proprie previsioni. Il nucleo della monografia è la dimostrazione formale che questa struttura — l'azione come riduzione dell'energia libera — implica necessariamente un modello causale del corpo nell'ambiente, e che i limiti di questo modello causale sono i limiti strutturali dell'agency biologica.

II. La struttura causale dell'azione volontaria

II.1 Il problema dell'intenzione e del controllo motorio

Il controllo motorio classico — la teoria delle servomeccanismi di Sherrington e la teoria del controllore ottimale di Bellman — descrive il movimento come la soluzione di un problema di ottimizzazione: dato un obiettivo (posizione finale dell'arto, forza da applicare, traiettoria da percorrere), trovare la sequenza di segnali motori che lo raggiunge minimizzando un costo (errore quadratico, energia muscolare, jerk — la derivata terza della posizione).

La formulazione classica del controllo ottimale è il problema di controllo di Hamilton-Jacobi-Bellman:

$$V(x, t) = \min_{u(\cdot)} \left\{ \int_t^T [\ell(x(s), u(s)) + \lambda \varphi(x(T))] ds \right.$$

dove $V(x, t)$ è la funzione valore (il costo minimo raggiungibile dallo stato x al tempo t), $\ell(x, u)$ è il costo istantaneo (running cost), $\varphi(x(T))$ è il costo terminale, $u(t)$ è il controllo (il segnale motorio), e la dinamica del sistema è:

$$\dot{x}(t) = f(x(t), u(t)) + w(t)$$

con $w(t) \sim N(0, Q)$ rumore di processo.

Il controllore ottimale lineare-quadratico (LQR) per sistemi con dinamica lineare ($f(x,u) = Ax + Bu$) e costo quadratico ($\ell(x,u) = x^T Qx + u^T R u$) produce la soluzione analitica:

$$u^*(t) = -K(t) \cdot \hat{x}(t)$$

dove $K(t)$ è il guadagno del controllore (soluzione dell'equazione di Riccati differenziale) e $\hat{x}(t)$ è la stima ottimale dello stato dal filtro di Kalman. L'ottimizzatore lineare-quadratico-gaussiano (LQG) combina il controllore LQR con il filtro di Kalman in modo disaccoppiato — il "separation principle" di Wonham (1968) garantisce che in sistemi lineari-gaussiani la stima ottimale (filtro di Kalman) e il controllo ottimale (LQR) si separano.

Ma il sistema nervoso non implementa un LQG — il sistema muscoloscheletrico è altamente non-lineare, i rumori sono non-gaussiani, e la struttura del compito cambia continuamente. Il modello di Todorov e Jordan (2002) — il "minimum intervention principle" — propone una versione più biologicamente realistica: il sistema di controllo motorio minimizza la varianza del risultato finale, non il costo del percorso, e interviene soltanto quando la deviazione dalla traiettoria intacca il risultato finale. Questo produce movimenti con alta variabilità nelle direzioni irrilevanti per il compito (bassa stiffness ortogonale) e bassa variabilità nelle direzioni rilevanti (alta stiffness parallela) — un pattern confermato sperimentalmente.

II.2 Il modello interno e la struttura causale del controllo

Il controllo motorio preciso richiede un "modello interno" del corpo — una rappresentazione della dinamica muscoloscheletrica che permette al sistema di predire le conseguenze sensoriali delle proprie azioni prima che queste accadano. Wolpert, Ghahramani e Jordan (1995) distinguono due tipi di modelli interni:

Il modello inverso $I: A \rightarrow M$ trasforma l'obiettivo dell'azione (la traiettoria desiderata A) nel comando motorio M necessario per produrre quella traiettoria. Il modello inverso è computazionalmente difficile perché la mappa da traiettoria a comando è altamente non-lineare e spesso mal condizionata (più comandi producono la stessa traiettoria — ridondanza muscolare).

Il modello diretto $D: M \rightarrow A$ trasforma il comando motorio M nella previsione della traiettoria A che ne risulterà. Il modello diretto è computazionalmente più semplice e serve a predire le conseguenze sensoriali delle azioni prima che i segnali di feedback arrivino (feedback delay problem).

Il sistema di controllo motorio biologico usa entrambi: il modello inverso per pianificare il movimento, il modello diretto per predire le conseguenze e generare segnali di errore predittivi. La struttura è quella di un controllore a feed-forward con feedback predittivo:

$$u(t) = I[A_desired(t)] + K \cdot [A_desired(t) - D[u(t-\tau)]]$$

dove il primo termine è il comando feed-forward dal modello inverso e il secondo termine è la correzione feedback basata sull'errore tra traiettoria desiderata e previsione del modello diretto (con ritardo τ per il tempo di calcolo).

III. L'inferenza attiva come unificazione di percezione e azione

III.1 La struttura del loop sensomotorio nel predictive coding

La Monografia VIII ha mostrato che nel framework del predictive coding, l'azione è la minimizzazione degli errori di previsione propriocettivi — le discrepanze tra la posizione attesa dell'arto (generata dal modello predittivo corticale) e la posizione attuale (misurata dai fusi neuromuscolari). L'azione "risolve" l'errore propriocettivo non aggiornando la previsione (come fa la percezione) ma cambiando lo stato corporeo per renderlo conforme alla previsione.

La formalizzazione precisa di questo meccanismo richiede la struttura del modello generativo per i segnali propriocettivi. Il modello generativo per l'arto in posizione θ con l'azione motoria u è:

$$\begin{aligned} \dot{s} &= f(\theta, u) + w_s \text{ (dinamica stato: l'arto si muove sotto il controllo } u) \\ y_p &= g(\theta) + v_p \text{ (segnale propriocettivo: il fuso neuromuscolare misura la posizione)} \end{aligned}$$

dove $s = \theta$ è lo stato (posizione dell'arto), f è la dinamica muscoloscheletrica, g è la funzione di risposta del fuso neuromuscolare, w_s e v_p sono rumori di processo e di osservazione. Il modello

predittivo del cervello mantiene la distribuzione posteriore $q(\theta)$ sulla posizione dell'arto.

L'errore di previsione propriocettivo è:

$$\epsilon_p = y_p - g(\mu_\theta)$$

dove $\mu_\theta = E_q[\theta]$ è la posizione attesa (il valor medio della distribuzione posteriore). La minimizzazione di F rispetto all'azione u dà:

$$du/dt = -\partial F/\partial u = -\partial F/\partial y_p \cdot \partial y_p/\partial u = \Pi_p \cdot \epsilon_p \cdot \partial g/\partial \theta \cdot \partial \theta/\partial u$$

dove Π_p è la precisione del segnale propriocettivo e $\partial \theta/\partial u$ è il guadagno motorio (quanto la posizione cambia per unità di comando motorio). L'azione si aggiorna nella direzione che riduce l'errore propriocettivo, pesato dalla precisione del segnale.

Questo meccanismo è implementato nel midollo spinale attraverso il riflesso gamma: i motoneuroni gamma innervano i fusi neuromuscolari e controllano la loro lunghezza di riferimento. Quando il cervello genera una previsione di posizione diversa dalla posizione attuale, il motoneurone gamma cambia la lunghezza di riferimento del fuso — generando un segnale di errore propriocettivo che attiva il riflesso miotattico (contrazione del muscolo agonista attraverso il motoneurone alfa). Il movimento è la conseguenza di questo riflesso: il corpo si muove per annullare l'errore propriocettivo creato intenzionalmente dalla variazione del segnale gamma.

III.2 L'azione come profezia che si autorealizza

Questa struttura — il sistema nervoso crea deliberatamente un segnale di errore propriocettivo attraverso la variazione del segnale gamma, e il corpo si muove per annullarlo — è la "profezia che si autorealizza" dell'azione volontaria: il cervello prevede uno stato futuro del corpo (la posizione desiderata dell'arto), questa previsione genera un errore rispetto allo stato attuale, e il corpo si muove per realizzare la previsione.

La struttura causale è la seguente (in notazione SCM della Monografia IX del ciclo precedente):

γ (segnale gamma, la previsione della posizione) $\rightarrow \epsilon_p$ (errore propriocettivo) $\rightarrow \alpha$ (attivazione motoneurone alfa) $\rightarrow M$ (contrazione muscolare) $\rightarrow \theta$ (posizione dell'arto) $\rightarrow y_p$ (feedback propriocettivo) $\rightarrow \epsilon_p$ (riduzione dell'errore)

Il loop causale è:

$\gamma \rightarrow [\epsilon_p = y_p - g(\mu, \theta)] \rightarrow \alpha \rightarrow M \rightarrow \theta \rightarrow y_p \rightarrow \epsilon_p$ (riduzione verso zero)

L'azione è quindi causalmente mediata dalla profezia (γ): il cervello genera una previsione della posizione desiderata attraverso il segnale gamma, e questa previsione causa il movimento. Non è il desiderio (una rappresentazione mentale) a causare il movimento attraverso un meccanismo misterioso; è la previsione incarnata nel segnale gamma a causare il movimento attraverso il riflesso propriocettivo — un meccanismo completamente fisico e computazionalmente preciso.

La struttura causale dell'azione volontaria è dunque:

$do(\gamma = \gamma_{target}) \rightarrow do(\theta = \theta_{target})$

dove l'intervento è sul segnale gamma (la previsione corticale della posizione) e la conseguenza è il cambiamento della posizione corporea. Questa è precisamente la struttura del $do(Alg = \pi)$ della Functional Decision Theory della Monografia XVII del ciclo precedente: l'intervento non è sull'azione diretta ma sul meccanismo (l'algoritmo, o in questo caso il segnale predittivo gamma) che genera l'azione.

IV. La causalità incarnata: il corpo come struttura causale

IV.1 Il concetto di affordance rivisitato

La Monografia III aveva analizzato il concetto gibsoniano di affordance come proprietà relazionale dell'ambiente rispetto alle capacità dell'agente. Il framework dell'inferenza attiva permette di riformulare questo concetto in termini di causalità: un'affordance è una struttura causale dell'ambiente che permette all'agente di ridurre la propria energia libera attraverso l'azione.

Formalmente, un'affordance è una relazione:

$Aff(o, a, F) \Leftrightarrow \exists u_t : \partial F / \partial u|_{u=u_t} < 0$ e l'azione u_t è resa possibile dalla relazione tra il corpo a e l'oggetto o

Un oggetto o offre un'affordance all'agente a se esiste un'azione u che riduce l'energia libera dell'agente nella presenza di o . Un gradino offre l'affordance di salire a un essere umano (l'azione di salire riduce l'energia libera rispetto all'obiettivo di arrivare più in alto) ma non a uno scarafaggio (le dimensioni relative non permettono l'azione).

Questa formalizzazione unifica la teoria gibsoniana delle affordance con il framework dell'inferenza attiva: le affordance non sono proprietà del mondo da rilevare passivamente, ma sono strutture causali dell'ambiente che rendono possibili certe riduzioni di energia libera. La percezione delle affordance è la rilevazione di queste strutture causali — non la costruzione di una rappresentazione del mondo, ma l'identificazione delle possibilità di azione che riducono la sorpresa.

La differenza rispetto alla versione della Monografia III è la connessione esplicita con la struttura causale: un'affordance non è soltanto la possibilità di un'azione, ma la struttura causale che rende quella azione efficace nel ridurre F. Questa connessione permette di derivare formalmente quali affordance un agente con un dato modello generativo del mondo percepirà — e quali errori di classificazione produrrà quando il modello generativo è mal calibrato.

IV.2 Il problema dei modelli del corpo: il rubber hand illusion

L'illusione della mano di gomma (Botvinick e Cohen, 1998) è un esperimento paradigmatico sulla struttura causale del modello corporeo. Il soggetto vede una mano di gomma essere toccata mentre il suo braccio reale (nascosto) viene toccato in modo sincronizzato nei medesimi punti. Dopo pochi minuti, il soggetto inizia a percepire la mano di gomma come parte del proprio corpo — e si aspetta che la mano di gomma senta il tocco — e mostra risposte galvaniche cutanee quando la mano di gomma viene minacciata con un coltello.

Nel framework del predictive coding, l'illusione si spiega come segue. Il modello generativo corticale del corpo comprende un prior sulla posizione del proprio arto. Quando il segnale visivo (vedere la mano di gomma toccare) è sincronizzato con il segnale tattile (sentire il braccio nascosto toccare), il sistema inferisce che la mano di gomma e il braccio nascosto appartengono allo stesso oggetto — il corpo dell'agente. Il prior sulla forma del corpo (braccia umane di certa forma e dimensione in certa posizione) ha alta precisione nella regione spaziale occupata dalla mano di gomma, e il segnale visivo-tattile sincronizzato aggiorna il posterior sulla posizione del corpo verso quella regione.

Formalmente, il modello generativo del corpo ha un prior:

$$p(\text{posizione_arto}) = N(\mu_body, \Sigma_body)$$

dove μ_body è la posizione media dell'arto e Σ_body è la covarianza che riflette l'incertezza sulla posizione. L'aggiornamento bayesiano con il segnale visivo-tattile sincronizzato produce:

$$p(\text{posizione_arto} \mid \text{visivo-tattile sincronizzato}) \propto p(\text{visivo-tattile} \mid \text{posizione_gomma}) \cdot p(\text{posizione_gomma})$$

con alta verosimiglianza per la posizione della mano di gomma (la sincronizzazione visuo-tattile è evidenza forte). Il posterior si sposta verso la posizione della mano di gomma — che viene incorporata nel modello corporeo.

La minaccia alla mano di gomma produce risposta galvanica perché il sistema predice segnali somatosensoriali negativi (dolore, danno corporeo) in risposta alla minaccia all'oggetto che ha incorporato nel proprio modello corporeo — un'altra profezia che si autorealizza, questa volta in direzione anticipatoria.

IV.3 I limiti del modello corporeo: la sindrome dell'arto fantasma

L'arto fantasma — la percezione persistente di un arto amputato — è il complemento patologico dell'illusione della mano di gomma: la mano di gomma mostra che il corpo può essere espanso per includere oggetti esterni; l'arto fantasma mostra che il corpo non può facilmente escludere regioni che hanno fatto parte di esso per decenni.

Nel framework del predictive coding, l'arto fantasma emerge perché il modello corporeo corticale include ancora la rappresentazione dell'arto amputato (il prior corporeo non viene aggiornato velocemente dopo l'amputazione) mentre i segnali propriocettivi dall'arto sono assenti. L'assenza di segnali propriocettivi dall'arto amputato — invece di aggiornare il modello corporeo verso uno stato senza quell'arto — produce errori di previsione persistenti (il modello prevede segnali che non arrivano) che il sistema interpreta come arto presente ma in stato anomalo (dolore fantasma, cramps, sensazioni strane).

Formalmente, il posterior sulla presenza dell'arto amputato è:

$$p(\text{arto_presente} \mid \text{assenza_segnale_propriocettivo}) \propto p(\text{assenza} \mid \text{arto_presente}) \cdot p(\text{arto_presente})$$

Il prior $p(\text{arto_presente})$ è molto alto (decenni di esperienza somatica dell'arto), e la verosimiglianza $p(\text{assenza} \mid \text{arto_presente})$ dipende dalla plausibilità che l'arto sia presente ma non produca segnali (possibile — un arto anestetico non produce segnali). Il posterior mantiene alta probabilità sulla presenza dell'arto anche in assenza di segnali — e l'arto "fantasma" viene percepito.

La soluzione terapeutica della mirror box di Ramachandran (1998) si inserisce in questa struttura: creare l'illusione visiva che l'arto amputato

stia muovendosi (riflettendo l'arto sano nello specchio) genera segnali visivi che, in sincronia con i comandi motori, aggiornano il modello corporeo verso uno stato in cui l'arto è presente e funzionale — riducendo il dolore fantasma attraverso la minimizzazione degli errori di previsione visivo-motori.

V. La causalità del corpo nell'ambiente: l'enattivismo formale

V.1 La tesi enattiva e la sua struttura matematica

Varela, Thompson e Rosch (1991) hanno proposto l'enattivismo come alternativa sia al cognitivismo (la mente come elaboratore di rappresentazioni) sia al connessionismo (la mente come pattern di attivazione di reti neurali): la cognizione è l'enazione di un mondo — la creazione co-emergente dell'agente e del suo ambiente attraverso il processo di vita stesso. Non vi è un mondo pre-dato che il sistema cognitivo rappresenta; vi è un dominio di azioni e percezioni che il sistema cognitivo fa emergere attraverso la propria struttura e il proprio accoppiamento strutturale con l'ambiente.

La formalizzazione matematica dell'enattivismo usa la biologia teorica di Maturana e Varela (1980) — l'autopoiesi. Un sistema autopoietico è un sistema che produce e mantiene continuamente la propria organizzazione attraverso le proprie operazioni:

Autopoiesi: M è autopoietico sse M produce i propri componenti attraverso reti di processi che costituiscono M , e questi componenti mantengono le reti di processi che li producono.

La struttura formale dell'autopoiesi può essere descritta come un'equazione differenziale sul manifold dei componenti del sistema. Sia C il set dei componenti del sistema e $P(C)$ il set dei processi di produzione che agiscono su C . La condizione di autopoiesi è:

$$\forall c \in C : \exists p \in P(C) : p \text{ produce } c \quad \forall p \in P(C) : i \text{ prodotti di } p \in C$$

Il sistema è chiuso rispetto alla produzione dei propri componenti. In un sistema autopoietico biologico (una cellula), C sono le molecole e le strutture cellulari, $P(C)$ sono le reazioni metaboliche e i processi anabolici.

L'enattivismo propone che la distinzione soggetto-oggetto — la distinzione tra l'agente che conosce e il mondo conosciuto — emerga

dall'autopoiesi stessa: il confine dell'organismo autopoietico (la membrana cellulare, la pelle) è al tempo stesso il confine epistemico — la divisione tra "interno" (il dominio delle rappresentazioni) e "esterno" (il mondo che il sistema incontra).

V.2 La struttura causale dell'accoppiamento strutturale

Il concetto centrale dell'enattivismo per la teoria dell'agency è l'accoppiamento strutturale: un sistema autopoietico e il suo ambiente si modificano reciprocamente attraverso le loro interazioni, producendo una storia di accoppiamenti che co-determina sia la struttura del sistema sia la struttura rilevante dell'ambiente.

Nel framework del predictive coding, l'accoppiamento strutturale corrisponde alla co-evoluzione del modello generativo (modificato dall'esperienza) e dell'ambiente (modificato dall'azione). L'agente modifica il proprio modello generativo per predire meglio l'ambiente (apprendimento); l'agente modifica l'ambiente attraverso le proprie azioni per renderlo conforme alle previsioni (inferenza attiva). L'accoppiamento strutturale è la traiettoria congiunta di questi due processi nel tempo.

Formalmente, sia q_t la distribuzione approssimante del sistema (il modello del mondo) e sia A_t la struttura rilevante dell'ambiente (le affordance percepite). La dinamica dell'accoppiamento strutturale è:

$$\begin{aligned} dq_t/dt &= -\eta \cdot \partial F(q_t, A_t)/\partial q \text{ (apprendimento: aggiorna il modello)} \\ dA_t/dt &= \varphi(u_t^*(q_t, A_t)) \text{ (azione: il sistema modifica l'ambiente)} \end{aligned}$$

dove u_t^* è l'azione ottimale sotto il modello q_t e φ è la funzione che descrive come l'azione cambia l'ambiente. Il sistema di equazioni differenziali accoppiate (q_t, A_t) descrive la co-evoluzione del sistema e del suo nicho epistemico.

Il punto fisso (q^*, A^*) di questo sistema — se esiste — è lo stato di equilibrio in cui il modello predice perfettamente l'ambiente e l'ambiente corrisponde perfettamente alle previsioni del modello. Questo è lo stato di perfetto accoppiamento strutturale — una forma di equilibrio ecologico tra l'agente e il suo nicho.

VI. I limiti della causalità incarnata

VI.1 Il problema dell'azione e della passività

La teoria dell'inferenza attiva descrive l'agente come un sistema che minimizza la propria energia libera attraverso l'azione. Ma ogni azione è anche una percezione di feedback — l'agente modifica l'ambiente e poi percepisce gli effetti di quella modifica. Questa struttura circolare produce un problema di causalità: chi è la causa e chi è l'effetto nell'interazione agente-ambiente?

La risposta del framework del predictive coding è che la distinzione causa-effetto è relativa alla struttura del modello generativo: il sistema tratta le proprie previsioni come cause (le previsioni top-down causano i segnali bottom-up attraverso la struttura gerarchica del modello) e i segnali sensoriali come effetti (le osservazioni sono gli effetti delle cause nel mondo). Ma questa distinzione è interna al modello — non è una proprietà del mondo fisico.

Il problema dell'azione e della passività si manifesta concretamente nell'analisi delle allucinazioni vs. percezioni: quando un agente "sente una voce" (allucinazione), il suo sistema cognitivo tratta il segnale come un effetto di una causa esterna (qualcuno sta parlando) invece che come un effetto della propria attività interna (pattern di attivazione corticale non causati da input sensoriale reale). La distinzione tra percezione (segnale causato dal mondo) e allucinazione (segnale causato dall'interno) non è determinata dal segnale stesso — che è identico nei due casi al livello neurale — ma dalla struttura del modello generativo che lo interpreta.

Formalmente, il sistema cognitivo assegna a ogni segnale di errore di previsione un'origine causale: le origini interne (generate dal modello stesso) producono aggiornamento delle previsioni (aggiornamento della credenza); le origini esterne (generate dal mondo) producono azione (cambiamento del mondo). Il meccanismo di questa distinzione — il segnale di "senso di agency" o "senso di autorità" (sense of agency) — dipende dalla coerenza tra il comando motorio prodotto dal sistema e il feedback sensoriale ricevuto.

VI.2 Il senso di agency e la struttura causale del sé

Il "senso di agency" — la sensazione che le proprie azioni siano causate dalla propria volontà — è un fenomeno studiato empiricamente da Haggard e collaboratori (Haggard e Eimer, 1999; Haggard et al., 2002) attraverso il paradigma del giudizio temporale. I soggetti producono un'azione volontaria (premere un tasto) e devono giudicare quando hanno preso la decisione di agire (sulla base della posizione di una lancetta rotante). Il potenziale di prontezza (Readiness Potential, RP) — un segnale EEG che precede l'azione di circa 800 ms — appare molto prima del momento soggettivo della decisione (circa 200–300 ms prima dell'azione secondo il giudizio soggettivo).

Questo risultato — il cervello "sa" che l'azione sta per avvenire 800 ms prima che il soggetto "decida" di agire — ha scatenato un dibattito decennale sul libero arbitrio. Nel framework del predictive coding, l'interpretazione è diversa: il momento soggettivo della decisione non è il momento in cui il cervello inizia a preparare l'azione, ma il momento in cui la preparazione dell'azione genera segnali di errore di previsione sufficientemente forti da penetrare la soglia della coscienza di accesso (dell'ignition fronto-parietale della GWT). Il RP è il segnale di preparazione sub-soglia; il momento soggettivo della decisione è la soglia di accesso cosciente.

Formalmente, sia $P_{\text{prep}}(t)$ il livello di preparazione motoria al tempo t (il segnale RP). L'azione è avviata quando $P_{\text{prep}}(t)$ supera la soglia motoria θ_{motor} . Il senso soggettivo di decisione avviene quando $P_{\text{prep}}(t)$ supera la soglia cognitiva $\theta_{\text{cognitive}}$. Tipicamente $\theta_{\text{cognitive}} < \theta_{\text{motor}}$ (l'accesso cosciente avviene dopo che la preparazione motoria ha raggiunto un certo livello, ma prima che l'azione sia iniziata):

$$t_{\text{RP}} \rightarrow t_{\text{cognitivo}} \rightarrow t_{\text{azione}}$$

con $t_{\text{RP}} \approx t_{\text{cognitivo}} - 600 \text{ ms}$ e $t_{\text{cognitivo}} \approx t_{\text{azione}} - 200 \text{ ms}$.

VII. Le niche costruite come prolungamento della causalità incarnata

VII.1 La teoria della costruzione della nicchia

Richard Lewontin (1983) propose che gli organismi non soltanto si adattano all'ambiente ma costruiscono attivamente il proprio ambiente — scelgono habitat, modificano il substrato, creano strutture fisiche, trasmettono questi cambiamenti alle generazioni successive. Odling-Smee, Laland e Feldman (1996, 2003) hanno formalizzato questa intuizione nella "niche construction theory" (NCT) come processo evolutivo distinto dalla selezione naturale.

Nel framework formale della NCT, l'organismo O costruisce una nicchia N attraverso una funzione di costruzione:

$$N_{t+1} = C(O_t, N_t, I_t)$$

dove C è la funzione di costruzione della nicchia, I_t sono i fattori ambientali indipendenti dall'organismo (perturbazioni esterne), e N_t è la nicchia al tempo t . La costruzione della nicchia produce feedback

evolutivi: la nicchia costruita modifica le pressioni selettive sull'organismo, che a sua volta modifica la nicchia costruita nelle generazioni successive.

Questo produce una co-evoluzione tra organismi e nicchie che ha struttura formalmente identica all'accoppiamento strutturale dell'enattivismo:

$dO_t/dt = f(O_t, N_t)$ (evoluzione degli organismi sotto le pressioni selettive della nicchia) $dN_t/dt = g(O_t, N_t)$ (costruzione della nicchia da parte degli organismi)

Il sistema di equazioni differenziali accoppiate (O_t, N_t) descrive la co-evoluzione organismi-nicchie, con traiettorie che dipendono dalle condizioni iniziali e dalla struttura delle funzioni f e g .

VII.2 Il linguaggio come costruzione della nicchia cognitiva

Il linguaggio è l'esempio più elaborato di costruzione di nicchia cognitiva nella specie umana. Attraverso il linguaggio, gli esseri umani creano un ambiente cognitivo condiviso — un "spazio semantico" in cui le informazioni possono essere trasmesse, manipolate e trasformate con precisione molto superiore a qualsiasi comunicazione pre-linguistica. Questo spazio semantico è esso stesso un ambiente che modifica le pressioni cognitive sugli individui che vi vivono: il linguaggio richiede l'elaborazione di strutture grammaticali complesse, la comprensione di riferimenti astratti, la manipolazione di proposizioni nidificate — capacità cognitive che si sviluppano soltanto nell'ambiente linguistico.

Kirby, Cornish e Smith (2008) hanno mostrato sperimentalmente come il linguaggio si auto-organizza per essere più trasmissibile: in esperimenti di "linguaggio iterated" (ogni generazione apprende una lingua artificiale da quella precedente), le lingue evolvono progressivamente verso strutture più composizionali e più regolari — strutture che riducono la complessità di Kolmogorov della lingua e quindi la rendono più facilmente trasmissibile. Il linguaggio naturale, in questa prospettiva, è il prodotto di migliaia di anni di selezione culturale per la trasmissibilità — è una nicchia cognitiva ottimizzata evolutivamente per essere facilmente acquisita e trasmessa.

La struttura formale è quella di un processo di selezione culturale sulla complessità del linguaggio:

$$K(L_{t+1}) = K(T(L_t)) + \delta$$

dove $K(L)$ è la complessità di Kolmogorov della lingua L , T è l'operatore di trasmissione (la funzione che trasforma la lingua della generazione t nella lingua appresa dalla generazione $t+1$), e $\delta > 0$ è la riduzione per generazione prodotta dalla pressione di trasmissibilità. Il linguaggio evolve verso la minima complessità compatibile con la necessità di esprimere l'insieme dei significati rilevanti per la cultura — un problema di ottimizzazione con vincoli che il processo evolutivo culturale risolve iterativamente.

VIII. I limiti dell'inferenza attiva come teoria dell'agency

VIII.1 Il problema della spiegazione del comportamento finalizzato

L'inferenza attiva descrive l'azione come minimizzazione degli errori di previsione rispetto a stati preferiti — uno "obiettivo" codificato come distribuzione prior fortemente concentrata sugli stati desiderati. Ma questo genera un circolo vizioso: per spiegare perché un agente persegue un dato obiettivo, il framework richiede che l'agente abbia già incorporato quell'obiettivo come prior — non spiega come gli obiettivi vengono formati o scelti.

Questo problema è formalmente identico al problema dell'indeterminazione degli obiettivi della Monografia XXIV del ciclo precedente: molti sistemi prior $p_{\text{target}}(x)$ sono compatibili con il comportamento osservato, e il framework del predictive coding non fornisce un criterio per selezionare tra essi. Un agente che minimizza F con $p_{\text{target}} = \text{"ottenere cibo"}$ e un agente che minimizza F con $p_{\text{target}} = \text{"muoversi verso il nord-est"}$ possono produrre comportamenti identici in certi ambienti.

La risposta del framework è che gli obiettivi si formano a loro volta attraverso la minimizzazione di F a un livello superiore della gerarchia — il livello che determina i prior di livello inferiore. Ma questo sposta il problema: come si formano gli obiettivi al livello superiore? Il regresso può essere fermato soltanto dall'evoluzione biologica (che ha selezionato i prior più in alto nella gerarchia) o dal condizionamento culturale (che ha selezionato i prior socialmente trasmessi). Il framework del predictive coding descrive come gli obiettivi vengono perseguiti, non come vengono formati — una limitazione strutturale che non può essere superata all'interno del framework stesso.

VIII.2 Il problema della creatività e dell'esplorazione

L'inferenza attiva è fondamentalmente un framework di minimizzazione: il sistema mira a stati di bassa energia libera, predice ciò che ha già modellato, agisce per realizzare previsioni già formate. Ma la creatività — la generazione di soluzioni, idee o comportamenti genuinamente nuovi — richiede l'esplorazione di stati non ancora previsti, la violazione intenzionale delle previsioni attuali.

Barto, Mirolli e Baldassarre (2013) e Oudeyer e Kaplan (2007) hanno proposto che la curiosità — il drive verso la novità e l'esplorazione — sia anch'essa derivabile dal principio di energia libera come minimizzazione della "free energy futura attesa" invece che della free energy attuale. Il sistema esplora stati non previsti perché la visita di quei stati ridurrà l'energia libera futura attraverso l'aggiornamento del modello generativo — un calcolo di long-term information value.

Formalmente, la curiosità è la preferenza per azioni a che massimizzano l'informazione attesa:

$$a^* = \operatorname{argmax}_a I(\theta; y_{t+1} \mid a, y_{1:t})$$

dove I è l'informazione mutua tra i parametri del modello θ e l'osservazione futura y_{t+1} . Questo è l'equivalente del "active learning" nell'apprendimento automatico: l'agente sceglie le osservazioni che massimizzano l'aggiornamento del proprio modello generativo.

Ma la creatività genuina — la generazione di soluzioni che non esistono come stati esplorabili nel modello corrente — rimane al di là della struttura dell'inferenza attiva. Un sistema che minimizza F (anche long-term F) può soltanto eseguire variazioni all'interno dello spazio di stati che il suo modello generativo considera possibili. La vera rottura dell'orizzonte del possibile — l'invenzione di un nuovo linguaggio, di una nuova tecnologia, di un nuovo sistema simbolico — richiede la generazione di stati che sono letteralmente impossibili nel modello corrente, il che richiede una metastruttura cognitiva al di sopra del modello che possa modificare il modello stesso. Questo meta-livello è, nel framework del predictive coding, parzialmente catturato dall'apprendimento gerarchico, ma la sua formalizzazione completa rimane un problema aperto.

IX. Conclusioni: il corpo come confine epistemico e causale

L'inferenza attiva, sviluppata in questa monografia, è il completamento necessario del principio di energia libera: mentre il predictive coding descrive come il sistema cognitivo costruisce previsioni e aggiorna le proprie credenze (la dimensione epistemica), l'inferenza attiva descrive come il sistema cognitivo agisce per realizzare le proprie previsioni (la dimensione causale). Il corpo è al tempo stesso il medium attraverso cui le previsioni vengono realizzate e lo strumento che vincola quali previsioni possono essere realizzate.

La causalità incarnata — la struttura causale che connette i comandi corticali ai movimenti corporei e i movimenti corporei al mondo — è la formalizzazione biologica di un'idea che filosofi come Merleau-Ponty avevano intuito senza poter formalizzare: il corpo non è un oggetto nel mondo ma la struttura attraverso cui il mondo si fa. Il segnale gamma che causa il movimento, il riflesso propriocettivo che realizza la previsione, l'accoppiamento strutturale che co-evolve agente e ambiente: questi sono i meccanismi fisici della "corporalità vissuta" di Merleau-Ponty, tradotti nel linguaggio del predictive coding.

I limiti di questa teoria — il regresso degli obiettivi, il problema della creatività genuina, la circolarità del modello generativo — non sono difetti occasionali ma confini strutturali: il framework descrive con precisione il territorio della cognizione incarnata stabilizzata, ma non cattura la giuntura tra il presente e il genuinamente nuovo. Comprendere questi confini è, come sempre in questo ciclo, il prerequisito per comprendere sia la struttura di ciò che è possibile sia la natura di ciò che rimane impossibile.

Monografia X

La connettività neurale come algebra

Grafi corticali, small-world networks, la struttura rich-club e la topologia funzionale dell'intelligenza

I. Il cervello come oggetto matematico

La neuroscienza del connettoma — la mappa completa delle connessioni neurali di un sistema nervoso — ha trasformato il cervello da oggetto biologico a oggetto matematico nel senso preciso in cui la fisica quantistica ha trasformato l'atomo: non ha eliminato la

complessità biologica, ma l'ha tradotta in strutture algebriche e topologiche che sono analizzabili con la strumentazione della matematica moderna. Il connettoma del nematode *Caenorhabditis elegans* — 302 neuroni e 7.000 sinapsi, completamente mappato da White et al. (1986) — è il prototipo di questa trasformazione: un grafo con 302 nodi e 7.000 archi orientati, studiabile con gli strumenti della teoria dei grafi, dell'algebra lineare e della topologia algebrica.

Il connettoma umano è enormemente più complesso: circa 86 miliardi di neuroni e 100–500 trilioni di sinapsi, equivalenti a un grafo con $\sim 10^{11}$ nodi e $\sim 10^{14}$ archi. La sua mappatura completa è attualmente impossibile — la risoluzione richiesta per rilevare ogni sinapsi in un volume di tessuto cerebrale dell'ordine dei mm^3 richiede decine di migliaia di terabyte. Il progetto Human Connectome (Van Essen et al., 2013) mappa invece il connettoma a livello macro (regioni corticali come nodi, fascetti di fibre come archi) attraverso la trattografia MRI diffusion: 80–180 regioni corticali connesse da centinaia di fascetti di fibre mielinizzate.

A questo livello di descrizione, il cervello è un grafo pesato $G = (V, E, W)$ con V le aree corticali (nodi), E le connessioni strutturali (archi) e W la matrice di pesi (forza delle connessioni). La domanda fondamentale di questa monografia è: quali proprietà algebriche e topologiche di questo grafo sono rilevanti per la funzione cognitiva, e quali strutture matematiche spiegano perché certi grafi neurali producono intelligenza flessibile mentre altri non la producono?

II. La teoria dei grafi applicata ai sistemi neurali

II.1 Proprietà fondamentali dei grafi neurali

La struttura di un grafo $G = (V, E)$ con $|V| = N$ nodi e $|E| = M$ archi è caratterizzata da diverse misure che catturano diverse dimensioni della sua organizzazione.

Il grado di un nodo v è $k_v = \#\{u : (u, v) \in E\}$ (per grafi non orientati) o la coppia (k_{in}^v, k_{out}^v) per grafi orientati. La distribuzione dei gradi $P(k)$ è la distribuzione della probabilità che un nodo scelto casualmente abbia grado k . Nei grafi neurali, $P(k)$ segue spesso una distribuzione a coda pesante (heavy-tailed) — una distribuzione in cui la probabilità di trovare nodi con grado molto alto decade come una legge di potenza $P(k) \propto k^{-\gamma}$ con $2 < \gamma < 3$.

La densità del grafo è $\rho = 2M / N(N-1)$ per grafi non orientati. I grafi neurali sono tipicamente sparsi: la corteccia cerebrale umana ha density $\rho \approx 0.1-0.3$ a livello macro, molto inferiore al grafo completo ($\rho = 1$) ma molto superiore a un grafo casuale con uguale numero di nodi ($\rho \approx k/N \approx 10^{-10}$ se ogni neurone avesse $k \approx 10.000$ connessioni su $N \approx 10^{11}$ neuroni).

Il coefficiente di clustering C_v del nodo v misura la frazione di vicini di v che sono anche connessi tra loro:

$$C_v = \#\{(u,w) \in E : u,w \in N(v)\} / (k_v(k_v-1)/2)$$

dove $N(v) = \{u : (u,v) \in E\}$ è il vicinato di v . Il clustering globale $C = (1/N) \sum_v C_v$ misura la tendenza del grafo a formare "triangoli" — pattern localmente clusterizzati.

Il cammino minimo L_{vw} tra due nodi v e w è il numero minimo di archi da attraversare per passare da v a w . La lunghezza media dei cammini minimi è:

$$L = 1/(N(N-1)) \sum_{v \neq w} L_{vw}$$

misura quanto "piccolo" è il mondo del grafo — quanto rapidamente l'informazione può propagarsi da qualsiasi nodo a qualsiasi altro.

II.2 Il modello small-world di Watts-Strogatz

Watts e Strogatz (1998) hanno proposto il modello "small-world" per caratterizzare le reti che combinano alta clusterizzazione locale con brevi cammini minimi globali — esattamente la struttura osservata nelle reti neurali, sociali e biologiche reali.

Il modello di Watts-Strogatz è costruito come segue: si parte da un grafo anello (lattice ring) in cui ogni nodo è connesso ai k vicini più prossimi su ciascun lato. Poi, con probabilità p , ogni arco viene "ricablato" (rewired) a un nodo scelto casualmente. Per $p = 0$, il grafo è un lattice regolare con alta clusterizzazione ($C \approx 3/4$ per k grande) e lunghi cammini minimi ($L \approx N/2k$). Per $p = 1$, il grafo è casuale con bassa clusterizzazione ($C \approx k/N \approx 0$) e brevi cammini minimi ($L \approx \log N / \log k$). Per p intermedio ($0 < p < 1$), il grafo mostra:

$$C(p) \approx C(0) \cdot f(p) \approx C(0) \text{ per piccoli } p \text{ (alta clusterizzazione preservata)}$$

$$L(p) \approx L_{\text{random}}(p) \approx \log N / \log k \text{ per piccoli } p \text{ (brevi cammini minimi)}$$

La transizione avviene in modo discontinuo: già per un piccolo numero di ricablature casuali ($p \approx 0.01-0.1$), i cammini minimi crollano verso i valori dei grafi casuali mentre la clusterizzazione rimane alta. Questo è il

"small-world effect": pochi collegamenti a lungo raggio (le cosiddette "weak ties" di Granovetter) sono sufficienti per ridurre drasticamente i cammini minimi senza distruggere la struttura clusterizzata locale.

La misura small-world σ di Humphries e Gurney (2008) cattura questa proprietà in modo quantitativo:

$$\sigma = (C/C_{\text{random}}) / (L/L_{\text{random}})$$

dove $C_{\text{random}} = k/N$ e $L_{\text{random}} = \log N / \log k$ sono la clusterizzazione e la lunghezza media dei cammini minimi per un grafo casuale con gli stessi N e k . Un grafo è "small-world" se $\sigma \gg 1$ — alta clusterizzazione normalizzata con breve lunghezza media normalizzata.

Per la corteccia cerebrale umana a livello macro: $\sigma \approx 2-3$ (Sporns e Zwi, 2004) — chiaramente nel regime small-world. Per il connettoma di C. elegans: $\sigma \approx 10$ (Watts e Strogatz, 1998) — fortemente small-world.

II.3 La giustificazione funzionale della struttura small-world

Perché i sistemi neurali sono small-world? La risposta di Sporns (2011) è una giustificazione funzionale: la struttura small-world bilancia due requisiti funzionali opposti del sistema cognitivo — la specializzazione locale (elaborazione modulare di informazione specifica di dominio) e l'integrazione globale (comunicazione rapida tra moduli specializzati).

La specializzazione locale è implementata dalla clusterizzazione: i nodi che elaborano informazione simile (ad esempio tutti i neuroni di V1 che rispondono a un certo orientamento) sono densamente connessi tra loro, formando moduli con alta coesione interna. L'integrazione globale è implementata dai brevi cammini minimi: qualsiasi due moduli possono comunicare in un piccolo numero di passi, permettendo l'integrazione dell'informazione su scala globale.

Il costo metabolico del cablaggio è il vincolo che forza questa struttura. Ogni connessione neurale richiede mielina, assoni, sinapsi — materiali biologicamente costosi. Un grafo completamente connesso ($q = 1$) avrebbe integrazione globale perfetta ma costo metabolico insostenibile. Un lattice regolare ($p = 0$) avrebbe minimo costo di cablaggio ma lunghissimi cammini minimi. La struttura small-world (p intermedio) è il punto di Pareto ottimale tra costo di cablaggio e prestazione cognitiva — il compromesso che l'evoluzione ha selezionato.

Questa è una formulazione precisa del principio di "accoppiamento ecologico" che percorre l'intero ciclo: l'architettura del grafo neurale è

accoppiata con i vincoli metabolici dell'ambiente corporeo e con i requisiti funzionali della cognizione adattiva.

III. La struttura rich-club e la gerarchia dell'integrazione

III.1 Il fenomeno rich-club e la sua formalizzazione

La struttura small-world cattura la topologia globale del grafo neurale, ma non ne cattura la struttura di gerarchia interna. Van den Heuvel e Sporns (2011) hanno identificato una proprietà ulteriore dei connettomi neurali: il "rich-club phenomenon" — la tendenza dei nodi ad alto grado (i "hub" della rete) a essere più densamente connessi tra loro di quanto ci si aspetterebbe da una distribuzione casuale dei link.

Il coefficiente rich-club $\varphi(k)$ misura la densità di connessioni tra i nodi con grado $\geq k$:

$$\varphi(k) = 2E_{\{>k\}} / N_{\{>k\}}(N_{\{>k\}} - 1)$$

dove $E_{\{>k\}}$ è il numero di archi tra nodi con grado $\geq k$ e $N_{\{>k\}}$ è il numero di nodi con grado $\geq k$. Il coefficiente normalizzato è:

$$\varphi_{\text{norm}}(k) = \varphi(k) / \varphi_{\text{rand}}(k)$$

dove $\varphi_{\text{rand}}(k)$ è il coefficiente atteso in un grafo casuale con la stessa distribuzione dei gradi. Un grafo mostra rich-club se $\varphi_{\text{norm}}(k) > 1$ per una certa soglia k — i nodi ad alto grado sono più connessi tra loro di quanto sia casuale.

Van den Heuvel e Sporns (2011) mostrano che la corteccia cerebrale umana ha $\varphi_{\text{norm}}(k) \approx 1.5\text{--}2.5$ per le soglie di grado rilevanti — mostrando forte rich-club. I nodi del rich-club corticale includono le aree precuneo, corteccia cingolata posteriore, corteccia prefrontale mediale, insula, talamo — aree che coincidono in larga parte con la Default Mode Network (DMN) e con il "default mode" di attivazione a riposo.

III.2 La struttura hub-and-spoke come architettura cognitiva

La struttura rich-club, combinata con la struttura small-world, produce un'architettura hub-and-spoke: hub altamente connessi tra loro (il rich-club) che fungono da centri di smistamento per l'informazione

proveniente dalle aree periferiche specializzate. Questa architettura ha una struttura algebrica precisa.

Sia A la matrice di adiacenza del grafo corticale ($A_{ij} = 1$ se esiste una connessione tra le aree i e j). Il vettore eigenvector centrality v è l'autovettore associato all'autovalore dominante λ_1 di A :

$$Av = \lambda_1 v$$

I componenti v_i dell'autovettore dominante sono la misura di centralità del nodo i : il nodo i è tanto più centrale quanto più i suoi vicini sono a loro volta centrali. Le aree del rich-club hanno alta eigenvector centrality — sono centrali perché i loro vicini (gli altri hub del rich-club) sono a loro volta centrali.

La comunicabilità del grafo — quanto facilmente due nodi possono comunicare — è misurata dalla matrice esponenziale:

$$\text{Comm} = \exp(\beta A) = \sum_{k=0}^{\infty} (\beta A)^k / k!$$

dove β è un parametro di temperatura inversa. Il termine $(\beta A)^k_{ij} / k!$ è il contributo dei cammini di lunghezza k da i a j . La comunicabilità tra i nodi i e j è:

$$\text{Comm}_{ij} = [\exp(\beta A)]_{ij}$$

che pesa i cammini di ogni lunghezza, con i cammini corti pesati più di quelli lunghi (il parametro β controlla questo trade-off). Le coppie di nodi nel rich-club hanno alta comunicabilità — possono scambiare informazione attraverso molteplici cammini brevi.

III.3 Il modello economico del connettoma

Bullmore e Sporns (2012) propongono di comprendere l'architettura del connettoma attraverso un framework di ottimizzazione economica: il cervello massimizza la prestazione cognitiva (connettività, efficienza di comunicazione) soggetto al vincolo del costo di cablaggio (volume fisico delle connessioni, energia metabolica per la trasmissione).

Il problema di ottimizzazione è:

$$\max_{\{G\}} \text{Perf}(G) \text{ soggetto a } \text{Cost}(G) \leq C_{\max}$$

dove $\text{Perf}(G)$ è la prestazione cognitiva (funzione della struttura topologica del grafo, ad esempio l'efficienza globale $E = (1/N(N-1)) \sum_{i,j} 1/L_{ij}$), $\text{Cost}(G)$ è il costo totale di cablaggio (somma delle distanze fisiche degli assoni pesate per i loro diametri), e $C_{\max}$ è il budget metabolico.

Il punto di ottimo di Pareto — il frontiera dei grafi che massimizzano la prestazione per un dato costo — ha la struttura small-world con rich-club: questa struttura è il compromesso ottimale tra costo di cablaggio e prestazione cognitiva nello spazio dei grafi sparsi.

La derivazione formale usa la relazione tra efficienza globale e struttura del grafo:

$$E(G) \approx E_{\text{lattice}}(G) + \alpha \cdot \varrho_{\text{longrange}}(G)$$

dove E_{lattice} è l'efficienza dovuta alle connessioni locali (alta clusterizzazione) e $\varrho_{\text{longrange}}$ è la densità di connessioni a lungo raggio (che riduce i cammini minimi). Per $\varrho_{\text{longrange}}$ piccolo, E cresce rapidamente con $\varrho_{\text{longrange}}$ — poche connessioni a lungo raggio producono grande aumento dell'efficienza (la "weak tie" di Granovetter). Il punto ottimale sotto il vincolo di costo è nel regime $\varrho_{\text{longrange}} \approx 0.01\text{--}0.1$ — esattamente il regime small-world.

IV. Reti di scale-free e la topologia dell'hub neurale

IV.1 Le reti scale-free e l'esponente di legge di potenza

Barabási e Albert (1999) hanno proposto che molte reti reali — internet, reti di citazioni, reti metaboliche, reti proteiche — abbiano distribuzioni dei gradi $P(k) \propto k^{-\gamma}$ (distribuzioni "senza scala" o scale-free). Questa distribuzione a legge di potenza emerge da un meccanismo di "preferential attachment" (attaccamento preferenziale): nuovi nodi si connettono preferenzialmente ai nodi già ad alto grado — "i ricchi diventano più ricchi".

Il meccanismo formale è il seguente. Sia G_t il grafo al passo t , e sia un nuovo nodo v_{t+1} aggiunto al passo $t+1$ con m nuovi archi. La probabilità che uno di questi archi si connetta al nodo j è proporzionale al grado corrente di j :

$$P(\text{arco verso } j) = k_j(t) / \sum_i k_i(t)$$

L'equazione di evoluzione per il grado atteso di j è:

$$d\langle k_j \rangle / dt = m \cdot \langle k_j \rangle / (2mt) = \langle k_j \rangle / (2t)$$

che dà la soluzione:

$$\langle k_j(t) \rangle = m \cdot (t/t_j)^{1/2}$$

dove t_j è il tempo in cui j è stato aggiunto al grafo. La distribuzione dei gradi nel limite $t \rightarrow \infty$ è:

$$P(k) = 2m^2 k^{-3}$$

una legge di potenza con esponente $\gamma = 3$.

Per i grafi neurali a livello macro, la distribuzione dei gradi non è esattamente una legge di potenza ma è heavy-tailed — la distribuzione è meglio descritta da una log-normale o da una legge di potenza troncata superiormente. L'esponente γ per la corteccia umana è stimato tra 2 e 3 da vari studi (Hagmann et al., 2008; van den Heuvel e Sporns, 2011), compatibile con il meccanismo di preferential attachment modificato da vincoli di costo di cablaggio.

IV.2 La robustezza e la vulnerabilità delle reti scale-free

Una proprietà topologica fondamentale delle reti scale-free è la loro "robustezza disomogenea": sono altamente robuste rispetto alla rimozione casuale di nodi (il fallimento casuale dei neuroni) ma altamente vulnerabili rispetto alla rimozione mirata degli hub (il danneggiamento selettivo delle aree hub del rich-club).

Albert, Jeong e Barabási (2000) hanno dimostrato formalmente questa proprietà. Per una rete scale-free con $P(k) \propto k^{-\gamma}$ e $2 < \gamma < 3$, la soglia di percolazione rispetto alla rimozione casuale di nodi è:

$$p_c^{\text{random}} = 1 - 1/(\kappa - 1) \text{ dove } \kappa = \langle k^2 \rangle / \langle k \rangle$$

Per distribuzioni a legge di potenza con $\gamma < 3$, $\langle k^2 \rangle$ diverge nel limite $N \rightarrow \infty$, il che implica $\kappa \rightarrow \infty$ e $p_c^{\text{random}} \rightarrow 0$. La rete rimane connessa anche dopo la rimozione di quasi tutti i nodi — infinita robustezza verso guasti casuali.

Per la rimozione mirata degli hub (rimozione dei nodi in ordine decrescente di grado), la soglia di percolazione è molto più bassa:

$$p_c^{\text{targeted}} \approx \exp(-1/\sqrt{\gamma-2}) \cdot (...)$$

che è dell'ordine di 0.1–0.2 per $\gamma \approx 2.5$ — bastano il 10–20% degli hub per distruggere la connettività della rete.

Applicato ai grafi neurali, questo risultato predice:

Le lesioni neurali casuali (infarti lacunari, neurotrasmissione aberrante diffusa) producono deficit cognitivi minimi fino a una soglia di danno molto alta — il cervello è robusto verso i guasti casuali.

Le lesioni mirate agli hub (tumori nelle aree del rich-club, atrofia preferenziale delle aree altamente connesse nelle malattie neurodegenerative) producono deficit cognitivi gravi con piccole lesioni anatomiche — il cervello è vulnerabile verso i guasti agli hub.

Questa predizione è confermata clinicamente: le malattie neurodegenerative (Alzheimer, Parkinson, sclerosi laterale amiotrofica) mostrano tipicamente atrofia preferenziale nelle aree del rich-club corticale — le aree più connesse sono anche le più vulnerabili, per ragioni metaboliche (alta domanda di energia per mantenere la connettività) e per ragioni topologiche (i prioni si diffondono preferenzialmente attraverso i hub, che sono i nodi più connessi).

V. La topologia algebrica applicata alle reti neurali

V.1 L'omologia persistente dei grafi neurali

La topologia algebrica — lo studio delle proprietà topologiche degli spazi matematici che sono invarianti rispetto alle deformazioni continue — offre strumenti per analizzare strutture del grafo neurale che sfuggono alla teoria dei grafi classica. In particolare, l'omologia persistente (Edelsbrunner e Harer, 2010) permette di identificare strutture multi-scala nel grafo — caratteristiche che persistono attraverso diverse scale di analisi.

Per un grafo pesato $G = (V, E, W)$, si costruisce una filtrazione: una sequenza di grafi G_ϵ anidati con ϵ che cresce da 0 a ∞ :

$$G_0 \subset G_{\{\epsilon_1\}} \subset G_{\{\epsilon_2\}} \subset \dots \subset G = G_\infty$$

dove G_ϵ include soltanto gli archi con peso $W_{\{ij\}} \geq \epsilon$. All'aumentare di ϵ (o equivalentemente, abbassando la soglia di connettività), archi vengono aggiunti al grafo e la topologia cambia.

Il complesso di Vietoris-Rips $\text{Rips}_\epsilon(G)$ è il complesso simpliciale costruito dal grafo G_ϵ : i vertici sono i nodi di G , gli archi sono i link di G_ϵ , i triangoli sono i triangoli di G_ϵ , i tetraedri sono i tetraedri, e così via. L'omologia $H_k(\text{Rips}_\epsilon)$ misura le strutture di dimensione k — i

"buchi" k -dimensionali — nel complesso simpliciale: H_0 conta le componenti connesse, H_1 i cicli, H_2 le cavità, e così via.

Il diagramma di persistenza traccia la nascita e la morte di queste strutture al variare di ϵ . Una caratteristica che nasce al valore $\epsilon_{\text{nascita}}$ e muore al valore ϵ_{morte} ha "lifetime" = $\epsilon_{\text{morte}} - \epsilon_{\text{nascita}}$. Le caratteristiche con lungo lifetime sono strutture topologiche "robuste" — persistono su un ampio intervallo di scale di analisi.

Giusti et al. (2015) hanno applicato l'omologia persistente al connettoma di *C. elegans* e hanno trovato che le strutture di H_1 (i cicli) e H_2 (le cavità) non compaiono in modo casuale ma si accumulano in regioni specifiche del grafo corrispondenti a circuiti funzionali noti — i circuiti motori, i circuiti sensoriali, i circuiti di integrazione. Questa è la prima evidenza che la topologia algebrica rivela strutture funzionalmente significative del connettoma che la teoria dei grafi standard non cattura.

V.2 La simplex neurale come unità computazionale

Reimann et al. (2017), nell'ambito del Blue Brain Project, hanno mostrato che la corteccia somatosensoriale del topo contiene un numero straordinariamente elevato di semplici completamente connessi (cliques): sottografi in cui ogni coppia di nodi è connessa da un arco. Il numero di cliques di dimensione k nel circuito corticale simulato è:

$$N(k) \approx N_0 \cdot r^k \text{ con } r \approx 1.5$$

che cresce (decrecendo più lentamente di un'esponenziale) fino a cliques di dimensione 7 o maggiore — semplici 6-dimensionali o superiori.

L'interpretazione funzionale proposta dal Blue Brain Project è che i semplici siano le unità computazionali fondamentali del circuito corticale: quando uno stimolo attiva uno simpleso (tutti i neuroni dello simpleso scaricano in sequenza), il sistema tratta questo schema di attivazione come un "pattern" coerente — equivalente a un "concetto" nell'architettura del modello generativo. Il numero di semplici e la loro dimensione determinano la "capacità rappresentazionale" del circuito — il numero di pattern distinguibili che il circuito può codificare.

La struttura algebrica sottostante è quella dell'algebra degli insiemi parzialmente ordinati (poset algebra): i semplici sono ordinati per inclusione (un simpleso di dimensione k è "sotto" ogni simpleso di dimensione $k+1$ che lo contiene), e le relazioni di inclusione formano un reticolo algebrico. Le operazioni di "join" (unione di semplici) e

"meet" (intersezione) corrispondono all'integrazione e alla segregazione dei pattern di attivazione.

VI. La struttura modulare del connettoma e la segregazione funzionale

VI.1 L'algoritmo di rilevamento delle comunità

La struttura modulare del grafo neurale — la tendenza delle aree corticali a formare moduli densamente connessi internamente e sparsamente connessi con l'esterno — è una delle proprietà topologiche più rilevanti per la funzione cognitiva. Il rilevamento delle comunità in grafi neurali usa tipicamente l'algoritmo di massimizzazione della modularità Q di Newman e Girvan (2004):

$$Q = (1/2M) \sum_{ij} [A_{ij} - k_i k_j / 2M] \delta(c_i, c_j)$$

dove A_{ij} sono i pesi degli archi, k_i è il grado del nodo i , $M = \sum_{ij} A_{ij} / 2$ è il numero totale di archi, c_i è la comunità a cui appartiene il nodo i , e $\delta(c_i, c_j) = 1$ se $c_i = c_j$. Il termine $A_{ij} - k_i k_j / 2M$ misura la connessione in eccesso rispetto a quella attesa in un grafo casuale con la stessa distribuzione dei gradi — la modularità misura quanto la struttura a comunità è più densa del caso.

Il problema di massimizzare Q su tutte le possibili partizioni è NP-hard (Brandes et al., 2008), ma algoritmi greedy (Louvain: Blondel et al., 2008) producono soluzioni sub-ottimali in tempo $O(N \log N)$ e identificano comunità stabili nei connettomi empirici.

Applicato al connettoma umano a livello macro, il rilevamento delle comunità identifica tipicamente 6–10 moduli stabili che corrispondono alle grandi reti funzionali note dall'fMRI: la rete visiva, la rete motoria, la rete di controllo esecutivo (frontoparietal network), la rete salience (insula e cingolato anteriore), la rete dorsal attention, la rete default mode (precuneo, mPFC, PCC). Queste reti mostrano alta modularità strutturale (alta densità di connessioni intra-modulo) e alta coattivazione funzionale (le aree nello stesso modulo strutturale tendono a attivare e disattivare insieme nell'fMRI a riposo).

VI.2 La flessibilità modulare e la plasticità cognitiva

La struttura modulare del connettoma non è statica: Bassett et al. (2011) hanno mostrato che la struttura modulare cambia nel corso dell'apprendimento. Usando la network science applicata all'fMRI

dinamico (analisi di serie temporali con finestra scorrevole), hanno mostrato che durante l'apprendimento di compiti visuomotori la struttura modulare del connettoma funzionale diventa più variabile — le comunità si riorganizzano più frequentemente — e questo aumento di flessibilità modulare è correlato con la velocità di apprendimento del compito.

La flessibilità modulare è formalizzata come:

$$\text{flex}(v) = (1/(T-1)) \sum_{t=1}^{T-1} \mathbb{1}[c_v(t) \neq c_v(t+1)]$$

dove $c_v(t)$ è la comunità del nodo v al tempo t e $\mathbb{1}$ è la funzione indicatrice. La flessibilità media del grafo è $\text{flex} = (1/N) \sum_v \text{flex}(v)$. Alta flessibilità indica che i nodi cambiano comunità frequentemente — il grafo funzionale si riorganizza spesso.

Bassett et al. (2011) trovano:

$\text{flex durante apprendimento} > \text{flex a riposo}$

e

$r(\text{flex_individuale}, \text{velocità_apprendimento}) > 0$ (correlazione positiva)

Questa correlazione tra flessibilità modulare e capacità di apprendimento ha un'interpretazione nel framework del predictive coding: la flessibilità modulare è la capacità del sistema di riconfigurare il proprio modello generativo — di formare nuove coalizioni di nodi che codificano nuove strutture causali. Alta flessibilità modulare corrisponde ad alta plasticità del modello generativo — capacità di aggiornare la struttura delle previsioni, non soltanto i valori dei parametri.

VII. Il connettoma come substrato della coscienza

VII.1 La struttura del connettoma e la teoria IIT

La teoria dell'informazione integrata (IIT, discussa nella Monografia IV) propone che la coscienza corrisponda al valore di Φ — l'informazione integrata del sistema. La struttura del connettoma determina il valore di Φ : sistemi con alta connettività a lungo raggio tra nodi diversi producono alto Φ ; sistemi con connettività puramente locale (come il cervelletto) producono basso Φ .

La relazione formale tra struttura topologica del connettoma e Φ può essere derivata per grafi semplici. Per un sistema binario a N nodi con matrice di connettività A , il valore di Φ è approssimativamente:

$$\Phi(G) \approx f(\lambda_1(A), \lambda_2(A), \dots, \lambda_N(A))$$

dove $\lambda_i(A)$ sono gli autovalori della matrice di adiacenza. Le reti con autovalori distribuiti su un ampio range (alta dispersione dello spettro) hanno alto Φ ; le reti con autovalori concentrati (come i grafi regolari) hanno basso Φ .

La distribuzione spettrale della corteccia cerebrale — misurata dalla decomposizione degli autovalori della matrice di connettività strutturale — mostra una distribuzione ampia e irregolare, con un grande autovalore dominante (corrispondente alla struttura rich-club globale) e molti autovalori intermedi (corrispondenti alle strutture modulari). Questa distribuzione produce alto Φ — in accordo con la previsione della IIT che la corteccia cerebrale, a differenza del cervelletto, ha alto livello di coscienza.

VII.2 Il problema dello scaling e la dimensione dell'integrazione

Un problema formale aperto nella IIT è lo scaling: il calcolo di Φ è NP-hard anche per sistemi moderatamente grandi (il numero di bipartizioni da considerare cresce come 2^N , e per $N \approx 80$ regioni corticali questo è astronomico). Approssimazioni computabili di Φ — come il "proxy phi" Φ^* e i metodi Monte Carlo — permettono stime per sistemi reali ma non il calcolo esatto.

Un problema più profondo è la scala di analisi: a quale livello del connettoma si deve calcolare Φ — a livello di neuroni individuali, di microcircuiti, di aree corticali, di reti? La IIT risponde che la coscienza è un fenomeno intrinseco al livello in cui Φ è massimo — il "maximum of integrated information" (MII). Ma identificare il MII in sistemi biologici reali è computazionalmente intrattabile con le tecnologie attuali.

VIII. I disturbi del connettoma come topologia patologica

VIII.1 La schizofrenia come disconnessione

La "disconnection hypothesis" della schizofrenia (Friston e Frith, 1995) propone che la schizofrenia sia causata da un'alterazione delle connessioni sinaptiche tra aree cerebrali — non da un danno localizzato a una singola area, ma da un'alterazione diffusa della struttura del grafo corticale. Le evidenze strutturali (riduzione della densità della materia bianca nelle connessioni fronto-temporali, fronto-parietali e fronto-talamiche) e funzionali (riduzione della connettività funzionale nelle reti frontali) supportano questa ipotesi.

La formalizzazione in termini di teoria dei grafi è la seguente. Il connettoma nella schizofrenia differisce dal connettoma normale in:

Riduzione della modularità Q : le comunità sono meno ben definite — la segregazione funzionale tra reti è ridotta.

Riduzione della small-worldness σ : il coefficiente di clustering è ridotto più di quanto si riduce la lunghezza media dei cammini minimi — il grafo è più "casuale" e meno "gerarchico".

Riduzione dell'eigenvector centrality degli hub: i nodi del rich-club hanno centralità ridotta — il "backbone" di integrazione globale è indebolito.

La conseguenza cognitiva di queste alterazioni topologiche — nel framework del predictive coding — è l'aberrazione della precisione descritta nella Monografia VIII: la riduzione della gerarchia hub-periferiche altera il calcolo della precisione dei segnali di errore, producendo un peso eccessivo sugli errori di previsione di basso livello (i segnali sensoriali "grezzi") rispetto alle previsioni di alto livello (i modelli del mondo).

IX. Conclusioni: il connettoma come linguaggio algebrico dell'intelligenza

Questa monografia ha mostrato che il connettoma — la struttura delle connessioni neurali — non è soltanto un substrato fisico dell'elaborazione cognitiva, ma è esso stesso un oggetto matematico con proprietà algebriche e topologiche che determinano le proprietà funzionali del sistema. La struttura small-world ottimizza il bilancio tra specializzazione locale e integrazione globale; il rich-club garantisce la comunicazione efficiente tra hub specializzati; la distribuzione scale-free dei gradi produce robustezza verso guasti casuali ma vulnerabilità verso lesioni agli hub; l'omologia persistente rivela strutture multi-scala che

corrispondono a circuiti funzionali; la struttura simpliciale determina la capacità rappresentazionale del circuito.

Il principio unificante è quello che percorre l'intero ciclo: l'architettura del connettoma è la soluzione biologica a problemi di ottimizzazione vincolata — massimizzare la prestazione cognitiva sotto il vincolo del costo metabolico e del costo fisico del cablaggio, in un ambiente evolutivo che ha selezionato questa architettura attraverso miliardi di generazioni. Il grafo neurale non è un oggetto arbitrario; è l'oggetto matematico che l'evoluzione ha costruito per risolvere il problema di elaborare informazione complessa in tempo reale con risorse metaboliche finite.

Comprendere questa struttura — e i limiti che essa impone alla prestazione cognitiva — è il prerequisito per comprendere sia l'intelligenza biologica nella sua grandezza sia le sue inevitabili impossibilità: quelle che derivano dalla finitezza del substrato, dalla topologia del grafo, e dai vincoli fisici che nessuna evoluzione può oltrepassare.

Monografia XI

La cognizione incarnata e il mind esteso

La tesi 4E, il corpo come substrato computazionale, la protesi cognitiva e i limiti della mente distribuita

I. Quattro aggettivi contro una tradizione

La tradizione cognitivista classica — che ha dominato le scienze cognitive dalla rivoluzione cognitivista degli anni Cinquanta fino alla fine del Novecento — concepisce la cognizione come elaborazione di simboli: il cervello riceve input sensoriali, li traduce in rappresentazioni interne, li manipola secondo regole formali, e produce output comportamentale. In questa visione, il corpo è il canale di input e di output, non il medium della cognizione; l'ambiente è il fornitore di stimoli, non il substrato dell'elaborazione; la cognizione è nel cervello, non nel corpo o nel mondo.

A partire dagli anni Novanta, questa visione è stata sistematicamente criticata da una famiglia di approcci raccolti sotto l'etichetta "4E cognition" — cognizione Embodied (incarnata), Embedded (situata),

Enacted (enattiva), Extended (estesa). Questi quattro aggettivi non sono sinonimi: indicano quattro tesi distinte con diversi gradi di radicalità e diversi gradi di supporto empirico. La tesi più moderata — la cognizione è incarnata — afferma che la struttura corporea modella la struttura cognitiva: non possiamo capire la cognizione senza capire il corpo che la implementa. La tesi più radicale — la cognizione si estende al di là del corpo — afferma che certi processi cognitivi sono letteralmente realizzati in parte fuori dal cervello e dal corpo, negli artefatti e negli ambienti che l'agente utilizza.

Questa monografia sviluppa le quattro tesi con la precisione formale che il programma del ciclo richiede, distinguendo con cura le tesi empiricamente supportate dalle congetture filosofiche, e analizzando i limiti strutturali di ciascuna. Il tema unificante è la domanda: dove finisce la mente? E la risposta che emerge è precisamente quella che ci si attende da un ciclo costruito intorno alle impossibilità strutturali — non una risposta netta, ma una mappa delle condizioni sotto cui la mente si estende e dei confini oltre i quali quella estensione fallisce.

II. La cognizione incarnata: il corpo come struttura cognitiva

II.1 Il cervello senza corpo: l'impossibilità del cervello in una vasca

L'esperimento mentale del "cervello in una vasca" — un cervello isolato dal corpo, mantenuto in vita artificialmente e connesso a un simulatore di input sensoriali — è stato usato da Putnam (1981) per argomentare contro il realismo semantico. Ma ha un'interpretazione neuroscientifica diretta: un cervello senza corpo non potrebbe funzionare normalmente, non soltanto per ragioni pratiche ma per ragioni strutturali.

La struttura del cervello è co-plasmata dalla struttura del corpo durante lo sviluppo — la Monografia V ha mostrato come i periodi critici calibrino le architetture corticali sull'input sensoriale corporeo. Ma la dipendenza va oltre il periodo critico: l'architettura corticale dell'adulto mantiene mappe del corpo (mappe somatotopiche nella corteccia motoria e sensoriale primaria) la cui struttura dipende dall'attività corporea corrente, non soltanto dallo sviluppo passato.

Le mappe somatotopiche (Penfield e Rasmussen, 1950) rappresentano le diverse parti del corpo in zone corticali diverse, con una topografia che riflette la rilevanza funzionale di ciascuna parte — il pollice e le

labbra hanno rappresentazioni corticali sproporzionatamente grandi rispetto alla loro dimensione fisica, riflettendo la loro importanza per la manipolazione fine e la comunicazione. Questa topografia non è statica: si riorganizza in risposta a cambiamenti nell'uso del corpo (musicisti che praticano strumenti a corda mostrano espansione delle rappresentazioni delle dita; soggetti che apprendono il Braille mostrano espansione delle rappresentazioni del dito indice sinistro).

La formalizzazione della dipendenza strutturale della corteccia dal corpo usa il framework della regolarizzazione adattiva (Adaptive Regularization):

$$W_{\text{cortex}}(t) = W_{\text{cortex}}(0) + \eta \cdot \int_0^t \frac{\partial \text{ELBO}}{\partial W} ds$$

dove W_{cortex} sono i pesi sinaptici corticali, ELBO è l'evidence lower bound del modello generativo corticale, e la derivata è calcolata rispetto all'esperienza sensorimotoria accumulata. La struttura $W_{\text{cortex}}(t)$ dipende dalla storia dell'esperienza corporea $\int_0^t x_{\text{body}}(s) ds$ — non soltanto dai dati sensoriali come tali, ma dalla struttura spaziotemporale dell'esperienza corporea.

Un cervello in una vasca, privato dell'input sensorimotorio strutturato, non evolverebbe semplicemente verso uno stato di riposo: degenererebbe progressivamente per mancanza del segnale di aggiornamento che mantiene la struttura delle mappe corticali. I sistemi di plasticità sinaptica richiedono attività per mantenersi — in assenza di attività, la plasticità omeostatica (Turrigiano, 2008) porta il sistema verso stati di ipereccitabilità che non corrispondono a nessuna elaborazione cognitiva funzionale.

II.2 La struttura del corpo come vincolo computazionale

Il principio di "morphological computation" (Pfeifer e Bongard, 2007) propone che la struttura fisica del corpo — la sua morfologia — computi parte dell'elaborazione che un sistema con corpo artificiale dovrebbe implementare esplicitamente nel controllo neurale. Il corpo non è soltanto il "hardware" su cui gira il "software" cognitivo — è parte del sistema computazionale.

L'esempio classico è la camminata: un sistema di cammino puramente controllato (un robot con attuatori rigidi e un controllore che specifica ogni grado di libertà) richiede enormi risorse computazionali per mantenere l'equilibrio e la fluidità del passo. Un sistema con morfologia biologicamente ispirata (gambe con elasticità passiva, tendini che immagazzinano energia) cammina in modo stabile con un controllore

molto più semplice — la morfologia "risolve" parte del problema di controllo passivamente, senza elaborazione neurale.

La formalizzazione usa il framework dell'informazione immagazzinata morfologicamente. Sia I_{body} l'informazione computazionale immagazzinata nella struttura fisica del corpo (morfologia, proprietà elastiche, inerzia) e sia I_{neural} l'informazione computazionale elaborata dal sistema nervoso. Il principio di morphological computation afferma:

$$I_{\text{total}} = I_{\text{body}} + I_{\text{neural}}$$

e la quantità $I_{\text{body}}/I_{\text{total}}$ — la frazione dell'elaborazione affidata alla morfologia — è una misura dell'efficienza morfologica del sistema. Per sistemi biologici reali (camminata bipedale umana), stime di $I_{\text{body}}/I_{\text{total}}$ sono dell'ordine di 0.3–0.5 (Pfeifer e Bongard, 2007): il 30–50% dell'elaborazione del movimento è affidata alla morfologia, non al sistema nervoso.

Questo ha implicazioni per la comprensione dell'evoluzione dei sistemi cognitivi: la selezione non opera soltanto sulla struttura del sistema nervoso, ma sulla struttura del corpo intero come sistema computazionale integrato. Un cambiamento morfologico (ad esempio l'evoluzione del piede umano con arco plantare elastico) riduce il requisito computazionale del sistema nervoso per il controllo del passo — un'innovazione cognitiva ottenuta attraverso un'innovazione morfologica, non neurologica.

III. La cognizione situata: l'ambiente come registro cognitivo

III.1 La distribuzione dell'informazione tra cervello e ambiente

Edwin Hutchins (1995) ha mostrato nella sua analisi della navigazione su navi militari che la cognizione non avviene in un cervello isolato che elabora rappresentazioni del mondo: avviene in un sistema distribuito che include cervelli, strumenti, procedure, artefatti e organizzazione sociale. La rete di strumenti nautici — compas, carte, quadranti, log — non è un'estensione del cervello ma un sistema cognitivo distribuito in cui il cervello è soltanto un componente.

Il caso della navigazione è particolarmente illuminante per la struttura formale della distribuzione cognitiva. La determinazione della posizione

di una nave richiede l'integrazione di misure di diversi strumenti (girobussola, log di velocità, radar, GPS), la correzione degli errori sistematici di ciascuno strumento, e il calcolo della posizione più probabile. Questo calcolo — che corrisponde a un filtro di Kalman nello stato posizione-velocità della nave — non avviene esclusivamente in nessun cervello individuale ma è distribuito tra:

- I marinai individuali che leggono i singoli strumenti e annotano le misure
- I protocolli di comunicazione che trasmettono le misure tra marinai
- I fogli di calcolo (o i computer) che eseguono la fusione delle misure
- Il capitano che integra le misure con la conoscenza delle condizioni locali

La struttura del filtro di Kalman è implementata dal sistema socio-tecnico, non da nessun cervello individuale. Hutchins chiama questo "cognition in the wild" — la cognizione nell'ambiente naturale delle pratiche umane, dove l'elaborazione è distribuita in strutture che nessuna teoria individualista della mente può catturare.

La formalizzazione usa la nozione di "computational complement" (complemento computazionale). Sia T un compito cognitivo con requisiti computazionali $C(T)$. Un agente con capacità cognitive $C_{\text{agent}} < C(T)$ non può completare il compito da solo. Ma se l'agente dispone di un complemento computazionale C_{artefact} nell'ambiente, il sistema integrato agente+artefatto ha capacità:

$$C_{\text{system}} = C_{\text{agent}} + C_{\text{artefact}} \geq C(T)$$

Il compito diventa eseguibile dall'agente con il complemento computazionale. Esempi: il calcolo della traiettoria balisticistica con una calcolatrice tascabile (C_{agent} basso, $C_{\text{calcolatrice}}$ alta, $C_{\text{system}} \geq C(T)$); la diagnosi medica con un sistema di supporto decisionale; la composizione musicale con uno spartito notato; la navigazione con un GPS.

La dipendenza dall'artefatto non è una debolezza cognitiva — è il design ottimale di un sistema cognitivo distribuito: esternalizzare l'elaborazione all'artefatto quando questo è più efficiente, e mantenere nell'agente biologico l'elaborazione che richiede le capacità uniche del cervello biologico (ragionamento contestuale, flessibilità, gestione dell'incertezza non strutturata).

III.2 La struttura delle scaffold cognitive

Vygotsky (1978) ha proposto la nozione di "zona di sviluppo prossimale" (ZPD) — la differenza tra ciò che un bambino può fare da solo e ciò che può fare con il supporto di un adulto. Il supporto dell'adulto — che Wood, Bruner e Ross (1976) chiamano "scaffolding" — è una struttura cognitiva temporanea che permette al bambino di eseguire compiti al di sopra della propria capacità attuale, e che viene progressivamente ritirata man mano che il bambino internalizza le strategie.

La struttura formale dello scaffolding cognitivo è quella della delega temporanea di elaborazione:

$$\text{Compito}(t) = F_{\text{bambino}}(t) \oplus F_{\text{adulto}}(t)$$

dove F_{bambino} è l'elaborazione del bambino, F_{adulto} è l'elaborazione dell'adulto (lo scaffold), e $\oplus$ è l'operazione di composizione che produce il risultato del compito. Nel corso dello sviluppo:

$$F_{\text{adulto}}(t) \rightarrow 0 \text{ mentre } F_{\text{bambino}}(t) \rightarrow F_{\text{competo}}(I)$$

Lo scaffold si ritira man mano che il bambino internalizza l'elaborazione. Il processo di internalizzazione di Vygotsky è la trasformazione della struttura computazionale distribuita (agente + scaffold) in una struttura computazionale interna (agente solo).

Questo processo ha un'implementazione neurale: l'internalizzazione corrisponde alla costruzione di un modello generativo che include la struttura dell'elaborazione precedentemente esternalizzata. Il bambino che prima usava le dita per contare, poi le immagini mentali delle dita, poi i simboli numerici — non perde l'accesso alle rappresentazioni esterne ma costruisce modelli generativi sempre più compatti che le sostituiscono funzionalmente.

IV. La mente estesa: Clark e Chalmers e il principio di parità

IV.1 La formulazione di Clark e Chalmers

Andy Clark e David Chalmers (1998) proposero la tesi della "mente estesa" nel loro saggio "The Extended Mind", uno dei più citati nella filosofia della mente contemporanea. La tesi è articolata attraverso il "principio di parità": se una parte del mondo funziona come parte della mente di un agente, e se rimuovere quella parte del mondo dall'agente

produrrebbe il medesimo deficit che produrrebbe rimuovere la corrispondente parte del cervello, allora quella parte del mondo è parte della mente dell'agente.

Il caso paradigmatico è Otto, un paziente con deficit di memoria episodica che usa sistematicamente un notebook per registrare le informazioni rilevanti sulla propria vita. Otto consulta il notebook automaticamente e con fiducia ogni volta che ha bisogno di informazione registrata — esattamente come Inga (con memoria normale) consulta automaticamente la propria memoria a lungo termine. Se la funzione di Otto+notebook è funzionalmente equivalente alla funzione di Inga con memoria normale, e se il notebook soddisfa le condizioni funzionali appropriate (prontamente disponibile, consultato abitualmente, accettato senza riserve), allora il notebook di Otto è parte del suo sistema cognitivo — il suo "sistema di credenze esteso" include le informazioni registrate nel notebook.

Clark e Chalmers formalizzano le condizioni per la "extended mind" come:

(1) Disponibilità: la risorsa esterna deve essere prontamente accessibile all'agente quando necessario. (2) Affidabilità: la risorsa deve essere sufficientemente affidabile — non produrre errori sistematici che l'agente non può rilevare. (3) Endorsement automatico: l'agente deve trattare le informazioni della risorsa come "proprie" senza riserve critiche sistematiche. (4) Accessibilità precedente: le informazioni devono essere state precedentemente "approvate" dall'agente come proprie.

Quando queste condizioni sono soddisfatte, la risorsa esterna funziona come parte del sistema cognitivo dell'agente — non metaforicamente, ma letteralmente nel senso funzionalmente rilevante.

IV.2 Le critiche formali alla mente estesa

La tesi di Clark e Chalmers ha generato un dibattito filosofico ampio. La critica più formalmente precisa è quella di Adams e Aizawa (2001, 2008): il "coupling-constitution fallacy" — la fallacia di confondere l'accoppiamento funzionale tra due sistemi con la costituzione dell'uno da parte dell'altro.

Adams e Aizawa argomentano che il fatto che due sistemi siano funzionalmente accoppiati (il cervello di Otto e il suo notebook producono insieme i risultati cognitivi desiderati) non implica che l'uno sia costituito (in parte) dall'altro. Il cervello di Otto e il suo notebook sono accoppiati ma distinti — esattamente come il cervello di un pilota

è accoppiato con il computer di bordo dell'aereo senza che il computer sia parte del cervello del pilota.

La critica formale può essere articolata in termini di causalità:

Se X è parte di Y, allora le relazioni causali interne a X sono interne a Y.

Se il notebook è parte del sistema cognitivo di Otto, allora le transizioni causali nelle pagine del notebook (inchiostro che asciuga, carta che si deteriora) sono transizioni causali all'interno del sistema cognitivo di Otto — una conclusione che sembra intuitivamente sbagliata.

Clark risponde che la causalità non è il criterio per la costituzione dei sistemi: ciò che conta è la funzione. I processi che servono la medesima funzione cognitiva (la registrazione e il recupero delle informazioni autobiografiche) sono parti del medesimo sistema cognitivo funzionale, indipendentemente dalla loro realizzazione fisica. La disputa si riduce a una disputa sulla natura dei sistemi cognitivi: strutturati dalla funzione o dalla causalità?

IV.3 La formalizzazione della mente estesa attraverso la teoria dei sistemi

Una formalizzazione più precisa della tesi della mente estesa usa la teoria dei sistemi (Bertalanffy, 1968) e la nozione di "sistema cognitivo" come sistema dinamico con frontiera funzionale invece che fisica.

Sia $S_{\text{cognitivo}}$ il sistema cognitivo di un agente, definito come il minimo insieme di stati che, insieme alla specificazione dell'ambiente, determinano il comportamento dell'agente nel dominio cognitivo di interesse. La frontiera di $S_{\text{cognitivo}}$ non è necessariamente la frontiera fisica del corpo: include tutti gli stati — corporei o extracorporei — che sono necessari per determinare il comportamento.

Formalmente, $S_{\text{cognitivo}}$ è il minimo chiusura causale del comportamento:

$$S_{\text{cognitivo}} = \min\{S : \forall \text{input } x, \forall t, \text{comportamento}(t) = f(S, x)\}$$

dove f è la funzione che mappa stati e input in comportamento. Se questa chiusura minimale include stati extracorporei (le informazioni nel notebook di Otto), allora la mente estesa è una tesi vera per quella chiusura.

La risposta di Adams e Aizawa in questi termini è che la chiusura causale del comportamento non è il criterio corretto per delimitare il sistema cognitivo: il criterio corretto è la chiusura intenzionale — la

chiusura dell'insieme degli stati che portano contenuto intenzionale (stati con aboutness, stati che rappresentano qualcosa). L'inchiostro nel notebook non porta contenuto intenzionale intrinseco — porta contenuto derivato (ha il significato che Otto gli ha assegnato), mentre gli stati cerebrali di Otto portano contenuto originale. Questa distinzione — tra contenuto intrinseco e contenuto derivato — è la linea di divisione tra mente corporea e artefatto extracorporeo.

V. La protesi cognitiva e i limiti dell'augmentation

V.1 La struttura funzionale delle protesi cognitive

Le protesi cognitive — tecnologie che aumentano le capacità cognitive di un agente al di sopra del normale — sono la versione tecnologica della mente estesa. Le calcolatrici tascabili, i sistemi GPS, i motori di ricerca, gli assistenti AI: tutti questi sistemi aumentano le capacità cognitive degli utenti estendendo funzionalmente le loro menti al di là dei limiti biologici.

La teoria delle protesi cognitive (Kirsh e Maglio, 1994) distingue tra azioni epistemic (azioni che modificano l'ambiente per facilitare l'elaborazione cognitiva) e azioni pragmatic (azioni che modificano l'ambiente per avanzare verso lo scopo). Un giocatore di Tetris che ruota i pezzi fisicamente per decidere se si adattano — invece di ruotarli mentalmente — compie un'azione epistemic: modifica l'ambiente (la posizione del pezzo) per facilitare un'elaborazione cognitiva (decidere se si adatta). Il costo fisico dell'azione epistemic è giustificato dalla riduzione del costo cognitivo dell'elaborazione mentale.

La struttura formale è quella dell'ottimizzazione del costo totale:

$$\min_{\{a_{\text{epistemic}}, a_{\text{pragmatic}}\}} [C_{\text{cognitive}}(a_{\text{epistemic}}) + C_{\text{physical}}(a_{\text{epistemic}}) + C_{\text{cognitive}}(a_{\text{pragmatic}}) + C_{\text{physical}}(a_{\text{pragmatic}})]$$

dove $C_{\text{cognitive}}$ è il costo di elaborazione neurale e C_{physical} è il costo di azione fisica. L'azione epistemic è ottimale quando il costo fisico dell'azione è inferiore alla riduzione del costo cognitivo che produce.

Kirsh e Maglio (1994) documentano che i giocatori esperti di Tetris compiono molte più rotazioni fisiche del necessario rispetto ai giocatori inesperti — una paradosso apparente che il framework dell'azione

epistemica spiega: i giocatori esperti hanno imparato che ruotare fisicamente riduce il costo cognitivo sufficientemente da giustificare il costo fisico aggiuntivo. Il calcolo ottimale del trade-off cognitivo-fisico è appreso, non dato a priori.

V.2 I limiti della protesi cognitiva: la dipendenza e la deskilling

La protesi cognitiva ha un lato oscuro che il framework funzionalista tende a sottostimare: la dipendenza e la perdita di competenza. Un agente che esternalizza sistematicamente una funzione cognitiva a una protesi può perdere progressivamente la capacità di eseguire quella funzione internamente — una forma di "cognitive offloading" (scarico cognitivo) che produce deskilling (perdita di competenza).

Il caso del GPS è paradigmatico. Ishikawa, Ramaprasad e Kastner (2008) hanno mostrato che l'uso sistematico di GPS produce riduzione della capacità di navigazione "dead reckoning" (navigazione per stima) e riduzione dell'attivazione dell'ippocampo durante compiti di navigazione. Tali effetti corrispondono formalmente alla riduzione della varianza delle distribuzioni posteriori sulla struttura spaziale dell'ambiente: chi usa il GPS smette di aggiornare la propria mappa cognitiva dell'ambiente, e il modello generativo spaziale si deteriora per mancanza di aggiornamento.

La formalizzazione usa il principio "use it or lose it" della neuroplasticità adulta. Sia W_{spatial} la matrice dei pesi corticali che codifica le competenze spaziali. La dinamica dei pesi nel regime di ridotta attivazione è:

$$dW_{\text{spatial}}/dt = \eta \cdot \partial \text{ELBO} / \partial W_{\text{spatial}} - \delta \cdot W_{\text{spatial}}$$

dove il primo termine è l'aggiornamento per attività (apprende quando il sistema è attivo) e il secondo termine è il decadimento omeostatico (i pesi non usati si indeboliscono). In assenza di attività (l'agente usa il GPS e non naviga autonomamente), il secondo termine domina:

$$W_{\text{spatial}}(t) \approx W_{\text{spatial}}(0) \cdot \exp(-\delta t)$$

Le competenze spaziali decadono esponenzialmente con la non-pratica. Questo produce una dipendenza strutturale dalla protesi: quanto più si usa il GPS, tanto meno si è capaci di navigare senza GPS, e tanto più si è motivati a usare il GPS — un loop positivo che produce dipendenza cognitiva.

Il limite formale della protesi cognitiva è quindi questo: la protesi aumenta la capacità cognitiva immediata ma può ridurre la capacità

cognitiva di riserva — la capacità di funzionare in assenza della protesi. In condizioni di disponibilità perfetta e affidabilità della protesi, questo trade-off è irrilevante. In condizioni di disponibilità incerta — temporanea perdita di segnale GPS, guasto della calcolatrice, mancanza di connessione internet — la dipendenza dalla protesi diventa una vulnerabilità.

VI. La cognizione distribuita nelle organizzazioni

VI.1 Le organizzazioni come sistemi cognitivi

Hutchins (1995) ha esteso l'analisi della cognizione distribuita dalle interazioni tra individuo e strumento alle organizzazioni come sistemi cognitivi. Un'organizzazione non è semplicemente una collezione di individui: è un sistema con proprietà cognitive emergenti che non appartengono a nessun individuo — memoria organizzativa (nelle procedure scritte, nelle tradizioni informali, nelle strutture fisiche), ragionamento distribuito (nelle procedure decisionali collettive), apprendimento organizzativo (nella capacità dell'organizzazione di modificare le proprie procedure in risposta all'esperienza).

La formalizzazione usa la teoria dei sistemi multi-agente. Un'organizzazione è un sistema $M = (A, S, C, T)$ dove $A = \{a_1, \dots, a\}$ sono gli agenti, $S = \{s_1, \dots, s\}$ sono le strutture (procedure, artefatti, gerarchie), $C: A \times A \rightarrow \{0,1\}$ è la matrice di comunicazione ($C_{ij} = 1$ se i comunica con j), e $T: (A \times S) \rightarrow \text{Outputs}$ è la funzione di trasformazione collettiva.

Le proprietà cognitive dell'organizzazione come sistema emergono dalla struttura (A, S, C, T) e non sono riducibili alle proprietà cognitive degli agenti individuali. La memoria organizzativa — la capacità dell'organizzazione di "ricordare" procedure e eventi passati — è distribuita nelle strutture S e non risiede in nessun agente individuale: quando tutti gli agenti che parteciparono a un evento storico lasciano l'organizzazione, la memoria dell'evento sopravvive nelle procedure scritte e nelle strutture fisiche.

VI.2 I limiti della cognizione organizzativa

La cognizione organizzativa ha limiti strutturali che corrispondono ai limiti della trasmissione culturale analizzati nella Monografia VII. La conoscenza tacita distribuita nelle pratiche organizzative — il "come si

fa" che non è scritto in nessun manuale — è soggetta alla stessa degenerazione per deriva della conoscenza tacita nelle culture tradizionali: quando le persone che portano quella conoscenza lasciano l'organizzazione, la conoscenza si perde, e nessun documento scritto può ricrearla completamente.

Il "turnover cognitivo" è il tasso a cui la conoscenza tacita dell'organizzazione viene persa per effetto del turnover del personale. Sia $K_{\text{tacita}}(t)$ la conoscenza tacita totale dell'organizzazione al tempo t e sia τ il tempo medio di permanenza degli agenti nell'organizzazione. La dinamica è:

$$dK_{\text{tacita}}/dt = -K_{\text{tacita}}/\tau + \text{Acquisizione_nuova}(t)$$

dove il primo termine è la perdita per turnover (ogni anno si perde $1/\tau$ della conoscenza tacita esistente) e il secondo termine è la nuova conoscenza tacita acquisita. All'equilibrio:

$$K_{\text{tacita}}^* = \tau \cdot \text{Acquisizione_nuova}$$

La conoscenza tacita dell'organizzazione è proporzionale al tempo medio di permanenza: organizzazioni con alto turnover (τ piccolo) hanno bassa conoscenza tacita e dipendono più fortemente dalla conoscenza esplicita codificata nelle procedure. Questo spiega perché le organizzazioni con alta rotazione del personale (alcune start-up, alcune agenzie governative) tendono a soffrire di "amnesia istituzionale" — la perdita sistematica della memoria delle ragioni delle procedure e delle lezioni apprese dagli errori passati.

VII. Il corpo esteso e i confini mobili del sé

VII.1 La plasticità del confine corporeo

La Monografia IX ha mostrato che il modello corporeo del cervello — il confine tra "sé" e "non-sé" — è plasticamente modificabile dall'esperienza. L'illusione della mano di gomma dimostra che oggetti esterni possono essere incorporati nel modello corporeo; la sindrome dell'arto fantasma dimostra che il modello corporeo può continuare a includere parti del corpo che non esistono più.

La plasticità del confine corporeo ha implicazioni per la tesi della mente estesa: se il modello corporeo è plasticamente modificabile, e se certi artefatti vengono incorporati nel modello corporeo (come avviene per il

bastone del cieco, per il violinista che "sente" il violino come parte del proprio corpo, per il chirurgo che usa il forcipe come estensione delle proprie dita), allora questi artefatti non soltanto funzionano come parte del sistema cognitivo dell'agente ma vengono letteralmente incorporati nella rappresentazione corticale del corpo.

Rizzolatti e Matelli (2003) hanno mostrato che i neuroni del sistema specchio nella corteccia premotoria dei primati rispondono non soltanto all'uso di utensili che prolungano il braccio, ma codificano la punta dello strumento come se fosse la mano stessa — il sistema sensorimotorio "ingrandisce" la mappa corporea per includere lo strumento. Questa non è una metafora: i campi recettivi tattili delle cellule bimodali (che rispondono sia a stimoli tattili sulla mano sia a stimoli visivi vicino alla mano) si espandono per includere lo strumento dopo alcune settimane di uso.

VII.2 I limiti dell'incorporazione: il rigetto del corpo esteso

L'incorporazione degli artefatti nel modello corporeo non è automatica e ha limiti strutturali. Per un artefatto esterno di essere incorporato nel modello corporeo corticale, devono essere soddisfatte alcune condizioni che corrispondono ai criteri di Clark e Chalmers per la mente estesa:

- (1) Consistenza sensorimotoria: il feedback sensoriale dall'artefatto deve essere coerente con le previsioni del modello sensorimotorio. Un guanto con feedback tattile ritardato di 100 ms produce sensazione di "straniamento" — il modello sensorimotorio predice il feedback tattile in sincronia con il movimento, e il ritardo produce un errore di previsione che impedisce l'incorporazione.
- (2) Continuità d'uso: l'artefatto deve essere usato con sufficiente continuità per permettere l'aggiornamento del modello corporeo. Uno strumento usato saltuariamente non viene incorporato; uno usato quotidianamente e per molte ore sì.
- (3) Adattabilità morfo-funzionale: l'artefatto deve essere adattato alla morfologia corporea — non ci si aspetta che una tastiera con layout AZERTY venga incorporata nel modello corporeo di un utente QWERTY che la usa occasionalmente.

Quando queste condizioni non sono soddisfatte, l'artefatto rimane un complemento computazionale esterno — un'estensione funzionale ma non un'estensione corporea. Il confine tra mente estesa (incorporazione funzionale) e cognizione distribuita (complemento esterno non incorporato) non è una dicotomia netta ma un continuum che dipende dal grado di integrazione sensorimotoria.

VIII. La linguistica cognitiva come prova della cognizione incarnata

VIII.1 La metafora concettuale come struttura corporea

Lakoff e Johnson (1999) hanno proposto, nella loro teoria della linguistica cognitiva, che la struttura del pensiero astratto umano sia radicalmente condizionata dalla struttura corporea — che i concetti astratti siano compresi e ragionati attraverso metafore primarie (primary metaphors) che mappano esperienze sensorimotoriche su domini astratti.

La metafora "AFFECTION IS WARMTH" (l'affetto è calore) emerge dall'esperienza corporea del calore fisico nella relazione madre-neonato: il calore corporeo della madre è associato all'affetto materno nell'esperienza preverbale del bambino. Questa associazione sensorimotorica diventa la base della metafora concettuale che struttura la comprensione adulta dell'affetto. Le espressioni linguistiche come "accoglienza calorosa", "persona fredda", "relazioni tiepide" non sono soltanto metafore linguistiche convenzionali — sono manifestazioni della struttura concettuale sottostante che radica l'affetto nel dominio sensorimotorico del calore corporeo.

La formalizzazione usa la struttura di un mapping tra domini:

φ : Dominio_sorgente (calore/freddo fisico) $\rightarrow$ Dominio_bersaglio (affetto/ostilità)

con corrispondenze strutturali:

calore fisico $\leftrightarrow$ affetto
freddo fisico $\leftrightarrow$ ostilità
aumento di temperatura $\leftrightarrow$ aumento dell'affetto ...

La mappatura è sistematica — non soltanto la coppia (calore, affetto) è mappata, ma l'intera struttura relazionale del dominio sorgente è proiettata sul dominio bersaglio. Questo è il "principio di invarianza" della teoria della metafora concettuale: la struttura del dominio sorgente è preservata nella mappatura.

La verifica sperimentale della teoria è sistematica. Williams e Bargh (2008) hanno mostrato che tenere in mano una tazza di caffè caldo (stimolo fisico) aumenta i giudizi di calore interpersonale di uno sconosciuto (giudizio sociale) — la metafora concettuale opera non soltanto a livello linguistico ma a livello di elaborazione percettiva.

Zhong e Leonardelli (2008) hanno mostrato che ricordare un'esperienza di esclusione sociale produce sensazione di freddo fisico — l'attivazione del concetto di "ostilità sociale" attiva il sistema neurale del freddo fisico, non soltanto la rappresentazione linguistica.

Questi risultati confermano che le metafore concettuali non sono soltanto strutture linguistiche ma strutture neurali che connettono sistemi sensorimotorici a sistemi di elaborazione astratta — la tesi dell'*embodied cognition* trova qui supporto sperimentale diretto.

IX. I limiti della cognizione 4E: dove la mente non si estende

La cognizione incarnata, situata, enattiva ed estesa è una descrizione accurata di molti aspetti dell'elaborazione cognitiva umana — la struttura del corpo modella la struttura cognitiva, l'ambiente fornisce risorse computazionali, l'azione co-costituisce la percezione, gli artefatti funzionano come parte del sistema cognitivo. Ma il framework ha limiti strutturali precisi che questa sezione enuncia.

Il primo limite è il problema dell'astrazione pura: certi domini cognitivi — la matematica avanzata, la fisica teorica, la logica formale — sembrano resistere alla riduzione alla struttura corporea. Un teorema di topologia algebrica non ha radici sensorimotoriche ovvie; il concetto di numero di Ramsey non è strutturato da nessuna metafora corporea primaria. Lakoff e Núñez (2000) hanno proposto che tutta la matematica sia radicata in metafore concettuali; ma la derivazione di certi risultati matematici astratti attraverso le metafore proposte è contestata.

Il secondo limite è il problema della universalità strutturale: se la cognizione è radicata nella struttura corporea, e le strutture corporee degli esseri umani sono simili ma non identiche tra culture, allora la struttura del pensiero dovrebbe variare sistematicamente tra culture. La ricerca cross-culturale sulla cognizione (Henrich, Heine e Norenzayan, 2010) documenta variazioni culturali significative in certi domini (percezione visiva, ragionamento causale, coerenza delle preferenze), ma mostra universali cognitivi forti in altri (struttura delle emozioni di base, percezione delle relazioni spaziali fondamentali, comprensione delle intenzioni altrui). La struttura corporea comune impone vincoli universali; le strutture corporee idiosincratiche permettono variazioni culturali.

Il terzo limite è il problema della coscienza: le tesi 4E descrivono la struttura funzionale della cognizione ma non affrontano il problema difficile della coscienza (Monografia IV). Un sistema robotico incarnato, situato e che esegue inferenza attiva realizza le tesi 4E funzionalmente, ma non si vede come questo implichi esperienza soggettiva. La coscienza rimane il confine irriducibile che nessuna teoria della cognizione incarnata ha attraversato.

X. Conclusioni: la mente come processo incarnato-situato-distribuito

Il percorso di questa monografia ha mostrato che la domanda "dove finisce la mente?" non ha una risposta univoca e netta — ha una risposta contestuale che dipende dalla funzione cognitiva considerata, dal grado di integrazione sensorimotoria con l'artefatto, dalla struttura dell'accoppiamento tra agente e ambiente.

Per le funzioni cognitive radicate nella struttura sensorimotorica del corpo (la percezione spaziale, la comprensione dell'azione, la comprensione dell'emozione attraverso le metafore corporee), il confine della mente coincide approssimativamente con il confine del corpo — ma non esattamente, perché il corpo si estende plasticamente per includere strumenti intensamente usati. Per le funzioni cognitive distribuite in sistemi socio-tecnici (la navigazione su navi, la produzione industriale, la ricerca scientifica), il confine funzionale della mente si estende al di là del corpo individuale per includere gli artefatti e le strutture organizzative che co-constituiscono l'elaborazione. Per le funzioni cognitive radicalmente astratte (la matematica, la logica formale), il confine della mente rimane ambiguo — c'è qualcosa che è strutturato corporalmente e qualcosa che sembra trascendere la struttura corporea.

Questa ambiguità non è un difetto della teoria — è la struttura reale del fenomeno. La mente non è un oggetto con confini netti; è un processo che si realizza attraverso cicli di azione-percezione che attraversano la frontiera corpo-ambiente in modo continuo e variabile. Comprendere la mente richiede comprendere questi cicli — non cercare un confine che non c'è.

Monografia XII

La termodinamica della vita

Strutture dissipative, neghentropia, il cervello come macchina termica e i limiti fisici della cognizione biologica

I. La vita contro la morte termica

La visione meccanicistica dell'universo che emerge dal XIX secolo porta in sé un paradosso. Il secondo principio della termodinamica — ogni sistema fisico isolato tende verso il massimo di entropia, verso il disordine, verso l'equilibrio termico — sembra incompatibile con il fatto ovvio dell'esistenza della vita: gli organismi viventi sono strutture altamente ordinate che si mantengono lontane dall'equilibrio per tutta la durata della loro vita, costruendo complessità invece di dissiparla. Come può la materia auto-organizzarsi in strutture complesse in un universo che tende verso il disordine?

Erwin Schrödinger, nel suo saggio "What is Life?" (1944), identificò la risposta fondamentale con una semplicità concettuale straordinaria: la vita si mantiene lontana dall'equilibrio non violando il secondo principio ma sfruttandolo — importando "ordine negativo" (neghentropia) dall'ambiente e scaricando entropia nell'ambiente. Un organismo vivente è una struttura che diminuisce localmente la propria entropia assorbendo energia libera dall'ambiente (cibo, luce solare) e dissipando calore nell'ambiente — mantenendo così il bilancio del secondo principio a livello del sistema totale (organismo + ambiente) mentre diminuisce localmente l'entropia interna.

Questa monografia sviluppa le implicazioni quantitative di questa visione per la biologia cognitiva: quali sono i limiti fisici imposti dalla termodinamica al funzionamento del cervello? Quanto energia richiede l'elaborazione cognitiva? Quanto lontano dal limite di Landauer opera il cervello biologico? E qual è la struttura termodinamica della cognizione — il rapporto tra la produzione di entropia cognitiva e il principio di energia libera della Monografia VIII?

II. La termodinamica del non-equilibrio e le strutture dissipative

II.1 Il teorema di Prigogine e le strutture lontane dall'equilibrio

Ilya Prigogine (Nobel per la Chimica, 1977) ha costruito la teoria delle strutture dissipative — sistemi fisici che mantengono organizzazione complessa attraverso il flusso continuo di energia e materia, dissipando entropia nell'ambiente. Le strutture dissipative sono la classe di fenomeni fisici a cui appartiene la vita: non sono strutture di equilibrio (che non evolvono) né strutture lontane dall'equilibrio che tendono al disordine (come i gas in espansione), ma strutture che si auto-organizzano e si mantengono stabili attraverso la dissipazione.

Il teorema del minimo di produzione di entropia di Prigogine (1945) stabilisce che nel regime lineare vicino all'equilibrio (dove i flussi termodinamici sono proporzionali alle forze termodinamiche), un sistema aperto a forze esterne costanti evolve verso uno stato stazionario che minimizza la produzione di entropia compatibilmente con i vincoli imposti dall'esterno:

$$d\sigma/dt \leq 0 \text{ per } t \rightarrow t_{\text{stazionario}}$$

dove σ è la produzione di entropia per unità di tempo e di volume:

$$\sigma = \sum_k J_k \cdot X_k$$

con J_k i flussi termodinamici (di calore, di materia, di carica elettrica) e X_k le forze termodinamiche (gradienti di temperatura, di concentrazione, di potenziale). Il teorema garantisce la stabilità degli stati stazionari vicino all'equilibrio — il sistema ritorna allo stato stazionario dopo perturbazioni.

Lontano dall'equilibrio (il regime biologico), il teorema di minimo non vale più. Le equazioni di evoluzione sono non lineari, e il sistema può evolversi verso stati stazionari multipli — stati con diversi livelli di organizzazione. Le transizioni tra questi stati — le biforcazioni — producono strutture dissipative: organizzazione spaziale spontanea (strutture di Turing nelle reazioni chimiche), oscillazioni temporali (cicli biologici), e strutture spaziotemporali complesse (i pattern di attivazione neurale nel cervello).

II.2 Le strutture di Turing e l'auto-organizzazione spaziale

Alan Turing (1952) propose un meccanismo matematico per la morfogenesi — la formazione di pattern spaziali negli organismi biologici — basato sull'instabilità diffusiva di sistemi di reazione-diffusione. Il sistema di Turing consiste di due sostanze

chimiche — un attivatore A e un inibitore I — che reagiscono tra loro e diffondono nello spazio con velocità diverse:

$$\frac{\partial A}{\partial t} = f(A, I) + D_A \nabla^2 A \quad \frac{\partial I}{\partial t} = g(A, I) + D_I \nabla^2 I$$

dove f e g sono le funzioni di reazione (tipicamente non lineari) e D_A , D_I sono i coefficienti di diffusione. La condizione di instabilità di Turing è:

$$D_I > D_A \text{ (l'inibitore diffonde più velocemente dell'attivatore)}$$

Quando questa condizione è soddisfatta, perturbazioni spaziali dello stato omogeneo si amplificano spontaneamente, producendo pattern spaziali periodici — "strisce" o "macchie" con lunghezza d'onda caratteristica:

$$\lambda_{\text{Turing}} = 2\pi/k_c \text{ dove } k_c = \sqrt{(\partial_A f \cdot D_I - \partial_I g \cdot D_A) / (D_A D_I)}$$

Il meccanismo di Turing è stato proposto per spiegare i pattern di colorazione della pelle degli animali (strisce della zebra, macchie del leopardo), la distribuzione dei follicoli piliferi, la formazione delle dita negli arti e i pattern di attivazione ritmica nei circuiti neurali (come le oscillazioni gamma nella corteccia).

Per la neuroscienza, il rilevante è che i pattern di Turing emergono spontaneamente dall'attività di sistemi di neuroni eccitatori e inibitori con range di connettività diversi — i neuroni eccitatori hanno connessioni locali (analogi all'attivatore), i neuroni inibitori hanno connessioni più estese (analogi all'inibitore). Questo schema "mexican hat" — eccitazione locale con inibizione laterale — è la struttura ricorrente dei circuiti corticali e produce pattern di attivazione oscillatoria nella banda gamma, beta e theta che corrispondono alle oscillazioni di Turing nel dominio spaziotemporale.

III. Il cervello come macchina termica

III.1 Il consumo energetico del cervello

Il cervello umano ha un metabolismo straordinariamente alto rispetto alla sua massa. Con circa 1.4 kg di massa (il 2% della massa corporea), il cervello consuma circa 20 W di potenza — il 20% del consumo energetico basale dell'intero corpo. Questa sproporzione indica un'intensità metabolica circa 10 volte superiore al metabolismo medio del resto del corpo.

Il consumo energetico del cervello è dominato dal mantenimento dei gradienti ionici attraverso le membrane neuronali — il lavoro della pompa Na^+/K^+ -ATPasi che mantiene il potenziale di membrana a riposo (-70 mV) ripristinando gli ioni che fuoriescono attraverso i canali di perdita. La potenza richiesta per mantenere il potenziale di membrana di un neurone è:

$$P_{\text{membrane}} = I_{\text{leak}} \cdot \Delta V$$

dove I_{leak} è la corrente di perdita attraverso i canali passivi e $\Delta V = 70 \text{ mV}$ è il potenziale di membrana. Per un neurone tipico con $I_{\text{leak}} \approx 0.1 \text{ nA}$ e $\Delta V = 70 \text{ mV}$:

$$P_{\text{membrane}} \approx 7 \times 10^{-12} \text{ W} = 7 \text{ pW per neurone}$$

Il consumo energetico totale di 86×10^9 neuroni è quindi:

$$P_{\text{totale}} \approx 86 \times 10^9 \times 7 \times 10^{-12} \text{ W} \approx 0.6 \text{ W}$$

Ma il cervello consuma 20 W — 33 volte di più. La differenza è dovuta alla trasmissione sinaptica: i potenziali d'azione richiedono la ripolarizzazione della membrana, ciascuna sinapsi richiede l'esocitosi di vescicole, e il ripristino delle concentrazioni ioniche richiede lavoro aggiuntivo della pompa Na^+/K^+ . Attergott et al. (1997) stimano che il 50% del consumo energetico sia dovuto alla trasmissione sinaptica, il 25% al mantenimento del potenziale di riposo, e il 25% a processi di supporto (sintesi delle proteine sinaptiche, trasporto assonale, mielina).

III.2 Il limite di Landauer applicato al cervello

Nella Monografia XXI del ciclo precedente abbiamo derivato il principio di Landauer: la cancellazione di 1 bit di informazione richiede la dissipazione di almeno $k_B T \ln 2 \approx 2.87 \times 10^{-21} \text{ J}$ a temperatura $T = 300 \text{ K}$. Il limite fisico assoluto per l'elaborazione cognitiva del cervello è quindi determinato dal numero di operazioni logiche per secondo che il cervello esegue.

Una stima conservativa del numero di operazioni sinaptiche per secondo nel cervello umano è:

$$N_{\text{operazioni}} \approx N_{\text{sinapsi}} \times f_{\text{media}} = 10^{14} \text{ sinapsi} \times 10 \text{ Hz} \approx 10^{15} \text{ operazioni/s}$$

dove ogni operazione sinaptica corrisponde approssimativamente a una operazione booleana (cancellazione di un bit di incertezza sul segnale). Il limite di Landauer per questo numero di operazioni è:

$$P_{\text{Landauer}} = N_{\text{operazioni}} \times k_B T \ln 2 = 10^{15} \times 2.87 \times 10^{-21} \text{ J/s} \approx 2.87 \times 10^{-6} \text{ W} \approx 3 \mu\text{W}$$

Il cervello consuma 20 W ma il limite di Landauer è 3 μW — il cervello è circa 7 milioni di volte meno efficiente del limite fisico assoluto. Questo sembra una critica all'efficienza del cervello, ma va contestualizzato: i transistor dei processori moderni operano a circa 10^6 volte il limite di Landauer — il cervello è circa 7 volte più efficiente dei processori moderni in termini di distanza dal limite di Landauer. La vicinanza al limite di Landauer riflette l'analogicità della computazione biologica e la bassa tensione di operazione delle membrane biologiche (70 mV vs. 1 V dei transistor).

La stima precisa del numero di bit cancellati per operazione sinaptica dipende dall'implementazione. Koch et al. (2006) stimano che un potenziale d'azione cancelli circa 10^4 bit di incertezza (la precisione temporale di un spike è circa 1 ms su un ciclo di ~ 100 ms, quindi $-\log_2(0.01) \approx 7$ bit, moltiplicato per il numero di sinapsi attivate $\approx 10^3$). Questo darebbe un costo per spike di circa $7000 \text{ bits} \times k_B T \ln 2 \approx 2 \times 10^{-17} \text{ J}$ per spike — ancora molto lontano dal limite di Landauer per la cancellazione fisica dei bit, ma coerente con il costo metabolico misurato ($\approx 10^{-15} \text{ J}$ per spike dalla misura del consumo di ATP).

III.3 L'efficienza termodinamica della cognizione

La cognizione biologica è termodinamicamente un processo di trasformazione di energia libera (glucosio) in lavoro informazionale (riduzione dell'incertezza sullo stato del mondo). L'efficienza di questa trasformazione può essere definita come:

$$\eta = I_{\text{prodotta}} / (\Delta G / k_B T \ln 2)$$

dove I_{prodotta} è l'informazione prodotta (in bit) e ΔG è l'energia libera consumata. Per il cervello umano:

$$I_{\text{prodotta}} \approx 10^{15} \text{ bits/s (dalla stima delle operazioni sinaptiche)} \quad \Delta G = 20 \text{ W} / (k_B T \ln 2) \approx 20 / (2.87 \times 10^{-21}) \approx 7 \times 10^{21} \text{ bits equivalenti/s}$$

$$\eta = 10^{15} / (7 \times 10^{21}) \approx 10^{-7}$$

L'efficienza termodinamica della cognizione biologica è circa 10^{-7} — un milionesimo dell'efficienza teorica massima. Questo valore sembra basso, ma riflette il fatto che la maggior parte dell'energia del cervello viene spesa per mantenere l'architettura neurale (i gradienti ionici, le strutture sinaptiche, i meccanismi molecolari) — non per l'elaborazione dell'informazione in senso stretto. L'energia "di mantenimento" è il costo fisso dell'hardware biologico.

IV. I cicli biologici come oscillatori termodinamici

IV.1 Il ritmo circadiano come orologio termodinamico

Il ritmo circadiano — il ciclo di circa 24 ore che sincronizza i processi biologici con il ciclo luce-buio — è un esempio paradigmatico di struttura dissipativa biologica. L'oscillazione circadiana non è una proprietà passiva del corpo (come un sistema che raffredda e si riscalda seguendo la temperatura ambientale), ma è un oscillatore attivo generato internamente che consuma energia per mantenere le oscillazioni.

Il meccanismo molecolare del ritmo circadiano nei mammiferi è basato su un loop di feedback trascrizionale-translazionale. I geni CLOCK e BMAL1 codificano per proteine che formano dimeri (CLOCK:BMAL1) e attivano la trascrizione dei geni Period (PER1-3) e Cryptochrome (CRY1-2). Le proteine PER e CRY si accumulano, formano complessi, inibiscono l'attività di CLOCK:BMAL1, e poi si degradano — permettendo un nuovo ciclo di attivazione. Il periodo del ciclo è determinato dai tassi di sintesi, accumulo e degradazione delle proteine.

La formalizzazione matematica del loop di feedback circadiano come oscillatore di Goodwin (1965) usa il sistema di equazioni differenziali:

$$\begin{aligned} dX/dt &= v_1 \cdot K^n / (K^n + Z^n) - v_2 \cdot X / (K_2 + X) & dY/dt &= k_3 \cdot X - v_4 \cdot Y \\ & / (K_4 + Y) & dZ/dt &= k_5 \cdot Y - v_6 \cdot Z / (K_6 + Z) \end{aligned}$$

dove X è la concentrazione dell'mRNA del gene Period, Y è la concentrazione della proteina PER non fosforilata, Z è la proteina PER fosforilata (che inibisce la trascrizione con cooperatività n), e i parametri $v_1, v_2, k_3, v_4, k_5, v_6$ sono i tassi di sintesi e degradazione. Il sistema ammette soluzioni oscillatorie (un ciclo limite stabile) per $n \geq 3$ (cooperatività sufficiente nell'inibizione).

Il ritmo circadiano è termodinamicamente un oscillo che consuma energia libera (ATP per la fosforilazione delle proteine, GTP per la sintesi proteica) per mantenere l'oscillazione. In assenza di input energetico, l'oscillazione si ammorza e si estingue — il ciclo non è passivo ma attivo. Il costo energetico del ritmo circadiano è stimato a circa 0.1–0.2 W per l'intero organismo — una frazione piccola del metabolismo basale ma non trascurabile.

IV.2 Le oscillazioni neurali come strutture dissipative

Le oscillazioni neurali — le oscillazioni sincrone dell'attività elettrica di grandi popolazioni di neuroni nella banda delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (15–30 Hz), gamma (30–80 Hz) — sono strutture dissipative che emergono dall'interazione non lineare tra popolazioni eccitatorie e inibitorie di neuroni. Come il ritmo circadiano, queste oscillazioni richiedono energia per essere mantenute e si estinguono in assenza di input energetico.

Il modello di Wilson-Cowan (1972) descrive la dinamica delle oscillazioni neurali come un sistema di equazioni differenziali non lineari per le popolazioni di neuroni eccitatori (E) e inibitori (I):

$$\tau_E \cdot dE/dt = -E + S_E(c_1E - c_2I + P_E) \quad \tau_I \cdot dI/dt = -I + S_I(c_3E - c_4I + P_I)$$

dove S_E e S_I sono le funzioni di risposta sigmoidali delle popolazioni, c_1 , c_2 , c_3 , c_4 sono le forze di connettività (eccitatoria e inibitoria), τ_E , τ_I sono le costanti di tempo, e P_E , P_I sono gli input esterni. Il sistema ammette un punto fisso stabile (attività costante), un ciclo limite stabile (oscillazione), o caos deterministico — a seconda dei parametri.

La transizione dal punto fisso al ciclo limite è una biforcazione di Hopf che avviene quando le connessioni eccitatorio-inibitorie hanno il rapporto appropriato. La frequenza del ciclo limite dipende dai parametri:

$$\omega_0 \approx \sqrt{(c_1c_3 - c_2c_4)} / (\tau_E \cdot \tau_I)$$

Per parametri corticali tipici, $\omega_0 \approx 30\text{--}70$ Hz — la banda gamma. La potenza consumata per mantenere l'oscillazione gamma è proporzionale all'ampiezza dell'oscillazione al quadrato:

$$P_{\text{gamma}} \propto A^2 \cdot \omega_0$$

dove A è l'ampiezza dell'oscillazione. Stime sperimentali danno $P_{\text{gamma}} \approx 10\text{--}30\%$ del consumo energetico corticale totale durante stati di alta attività cognitiva (attenzione sostenuta, elaborazione visiva complessa).

V. La termodinamica dell'apprendimento

V.1 Il costo energetico della potenziale a lungo termine

La potenziale a lungo termine (LTP) — il meccanismo cellulare fondamentale della memoria sinaptica — richiede un costo energetico significativo. L'LTP coinvolge: l'inserzione di nuovi recettori AMPA nella membrana postsinaptica (richiede sintesi e trasporto proteico), la modificazione della struttura delle spine dendritiche (richiede polimerizzazione dell'actina), la sintesi di nuove proteine (richiede ribosomi e ATP), e la fosforilazione di CaMKII (richiede ATP).

Il costo energetico dell'LTP per singola sinapsi è stimato a circa 10^4 – 10^5 molecole di ATP, corrispondente a circa 10^{-18} – 10^{-17} J per sinapsi. Per un cervello che forma 10^4 nuovi ricordi al giorno con aggiornamento di $\sim 10^3$ sinapsi per ricordo, il costo energetico totale dell'apprendimento è:

$$P_{\text{LTP}} \approx 10^4 \text{ ricordi/giorno} \times 10^3 \text{ sinapsi/ricordo} \times 10^{-17} \text{ J/sinapsi} / (86400 \text{ s/giorno})$$

$$P_{\text{LTP}} \approx 10^{-12} \text{ W} = 1 \text{ pW}$$

Questo è trascurabile rispetto al consumo energetico totale del cervello (20 W) — l'apprendimento è metabolicamente "economico" rispetto alla manutenzione dell'architettura neurale. Questo risultato sembra sorprendente ma è coerente con l'osservazione che il cervello umano può apprendere continuamente per tutta la vita senza apprezzabili aumenti del consumo energetico.

La conseguenza termodinamica è importante: il vincolo energetico non è il fattore limitante dell'apprendimento biologico. I fattori limitanti sono di tipo diverso — il numero di sinapsi disponibili (la capacità di stoccaggio), la velocità di consolidazione (che richiede sonno e riposo), e l'interferenza tra vecchi e nuovi ricordi (il catastrophic forgetting biologico, mitigato dalla struttura complementare dell'apprendimento ippocampale-corticale).

V.2 La termodinamica della consolidazione della memoria

La consolidazione della memoria — il processo attraverso cui i ricordi episodici recenti (ippocampali) vengono gradualmente trasferiti a rappresentazioni corticali stabili — ha una struttura termodinamica precisa. Tononi e Cirelli (2003, 2006) hanno proposto la "synaptic homeostasis hypothesis" (SHY): durante la veglia, l'apprendimento produce un netto aumento del peso sinaptico medio nella corteccia; durante il sonno, la potenza sinaptica viene ridotta verso un livello basale attraverso la depressione a lungo termine (LTD) — un processo di "normalizzazione" che riduce il rumore sinaptico e selettivamente rafforza le sinapsi più importanti.

La SHY ha una struttura termodinamica elegante. Durante la veglia:

$$\sum_{\text{sinapsi}} W_{\text{sinapsi}}(t) > \sum_{\text{sinapsi}} W_{\text{sinapsi}}(0)$$

L'apprendimento aumenta la somma totale dei pesi sinaptici. Durante il sonno:

$$\sum_{\text{sinapsi}} W_{\text{sinapsi}}(t + \text{notte}) < \sum_{\text{sinapsi}} W_{\text{sinapsi}}(t)$$

Il sonno riduce la somma totale dei pesi sinaptici verso il livello basale. La differenza:

$$\Delta W_{\text{netto}} = \sum W_{\text{sinapsi}}(t) - \sum W_{\text{sinapsi}}(0) \approx 0 \text{ su scale di tempo giorni}$$

Il cervello è in omeostasi sinaptica su scale di tempo di giorni: cresce durante la veglia (apprendimento), scende durante il sonno (normalizzazione).

La termodinamica della SHY è quella di un sistema in stato stazionario fuori dall'equilibrio: il flusso di energia cognitiva (apprendimento) aumenta l'entropia sinaptica (complessità della rete sinaptica), e il sonno riduce l'entropia sinaptica attraverso un processo selettivo (LTD delle sinapsi deboli, mantenimento delle sinapsi forti). Il costo energetico del sonno — circa 1–2 W in meno rispetto alla veglia (il cervello addormentato consuma meno) — è il costo della normalizzazione sinaptica.

VI. Le leggi di scala metabolica e il cervello

VI.1 La legge di Kleiber e le leggi di potenza del metabolismo

Kleiber (1932) ha scoperto che il metabolismo basale degli organismi scala come una legge di potenza della massa corporea:

$$B \propto M^{3/4}$$

dove B è il metabolismo basale (in watt) e M è la massa corporea (in kg). Questa "legge di Kleiber" è stata successivamente derivata da West, Brown e Enquist (1997) come conseguenza delle proprietà di ottimizzazione delle reti vascolari che distribuiscono ossigeno e nutrienti alle cellule — reti che si strutturano come frattali gerarchici per minimizzare la resistenza alla distribuzione.

La derivazione di West et al. (1997) procede come segue. Sia la rete vascolare un albero frattale con C livelli, in cui il flusso sanguigno si ramifica a ciascun livello di un fattore n (ogni vaso si divide in n vasi più piccoli). La massa corporea M è proporzionale al volume totale dell'organismo, e il metabolismo B è proporzionale alla velocità totale di trasporto dell'ossigeno ai tessuti. La minimizzazione della resistenza idrodinamica soggetta al vincolo di area terminale costante (i capillari hanno dimensioni fisse) dà:

$$B \propto M^\alpha \text{ con } \alpha = 1 - 1/C_{\text{dim}}$$

dove C_{dim} è la dimensione frattale della rete. Per reti tridimensionali con ramificazione ottimale, $C_{\text{dim}} = 4$ e $\alpha = 3/4$ — recuperando la legge di Kleiber.

L'implicazione per il cervello è che il metabolismo cerebrale scala con la massa corporea secondo:

$$B_{\text{cervello}} \propto M^{(3/4)}$$

ma la massa cerebrale scala più lentamente:

$$M_{\text{cervello}} \propto M^{(3/4)} \text{ anche}$$

Il rapporto $B_{\text{cervello}}/M_{\text{cervello}}$ (il metabolismo cerebrale per unità di massa) rimane approssimativamente costante attraverso le specie — circa 11 W/kg per la maggior parte dei mammiferi. L'eccezione sono i primati (specialmente i primati umani) in cui il rapporto $B_{\text{cervello}}/B_{\text{totale}}$ è eccezionalmente alto: il 20–25% per gli umani contro il 2–5% per la maggior parte dei mammiferi. Questo suggerisce che l'evoluzione del cervello grande negli ominidi ha richiesto una riorganizzazione del metabolismo corporeo per sostenere il costo energetico dell'encefalizzazione.

VII. La termodinamica dell'evoluzione del cervello umano

VII.1 Il costo di Aiello e Wheeler

Aiello e Wheeler (1995) hanno proposto la "expensive tissue hypothesis" per spiegare l'evoluzione del grande cervello negli ominidi: il cervello è metabolicamente "costoso" (alto metabolismo per unità di massa) e il suo aumento di dimensione nei nostri antenati è stato possibile soltanto attraverso una riduzione compensatoria di altri tessuti

metabolicamente costosi. Il candidato principale è il tratto gastrointestinale (intestino): nell'evoluzione degli ominidi, il grande cervello si è sviluppato di pari passo con un intestino ridotto — coerente con una dieta di alta qualità (ricca di nutrienti facilmente digeribili, come carne e tuberi cotti) che permette di estrarre le stesse calorie con un intestino più piccolo.

La formalizzazione è quella di un'ottimizzazione vincolata. Sia E_{cervello} il costo energetico del cervello, $E_{\text{intestino}}$ il costo del tratto gastrointestinale, e E_{totale} il budget energetico totale disponibile. Il vincolo è:

$$E_{\text{cervello}} + E_{\text{intestino}} + E_{\text{altri}} \leq E_{\text{totale}}$$

Se E_{totale} è fisso (determinato dalla dieta), e se E_{cervello} aumenta (evoluzione verso cervello più grande), allora almeno un termine degli altri deve diminuire. La proposta di Aiello e Wheeler è:

$$\Delta E_{\text{cervello}} = -\Delta E_{\text{intestino}} + \epsilon$$

dove ϵ è un termine piccolo — l'aumento del cervello è quasi esattamente compensato dalla riduzione dell'intestino, con un piccolo guadagno netto reso possibile dalla dieta di alta qualità.

Questa ipotesi è supportata da misure comparative delle dimensioni di cervello e intestino in primati di diverse dimensioni: la correlazione negativa tra dimensione cerebrale e dimensione intestinale, controllando per massa corporea, è statisticamente significativa ($r \approx -0.7$, $p < 0.001$, Aiello e Wheeler 1995).

VII.2 Il costo del sonno come necessità termodinamica

Se la SHY di Tononi e Cirelli è corretta, il sonno non è un lusso evolutivo — è una necessità termodinamica del cervello grande. Un cervello con molte sinapsi (come il cervello umano) ha un tasso di accumulo sinaptico durante la veglia (per effetto dell'apprendimento) più alto di un cervello piccolo. Senza sonno, il cervello raggiungerebbe una saturazione sinaptica — uno stato in cui tutte le sinapsi sono al massimo peso e non possono più aumentare — impedendo ulteriore apprendimento.

La formalizzazione usa la capacità sinaptica totale C_{max} e il tasso di accumulo sinaptico durante la veglia r_{veglia} :

$$dW_{\text{totale}}/dt = r_{\text{veglia}} \text{ durante la veglia} \quad dW_{\text{totale}}/dt = -r_{\text{sonno}} \text{ durante il sonno}$$

All'equilibrio, se T_{veglia} è il tempo di veglia e T_{sonno} è il tempo di sonno:

$$r_{\text{veglia}} \cdot T_{\text{veglia}} = r_{\text{sonno}} \cdot T_{\text{sonno}}$$

Il rapporto $T_{\text{veglia}}/T_{\text{sonno}} = r_{\text{sonno}}/r_{\text{veglia}}$ deve essere tale da mantenere W_{totale} al di sotto della saturazione C_{max} . Per cervelli più grandi (con più sinapsi), C_{max} è maggiore ma r_{veglia} scala con il numero di sinapsi attivate per unità di tempo — che può essere più alto nei cervelli più complessi. La necessità di sonno per la normalizzazione sinaptica è una conseguenza termodinamica della dimensione del cervello.

VIII. Le strutture dissipative cognitive: creatività e insight

VIII.1 L'insight come biforcazione termodinamica

I fenomeni di insight — la comprensione improvvisa di una soluzione a un problema ("aha moment") — possono essere interpretati nel framework delle strutture dissipative come biforcazioni del sistema cognitivo: transizioni di fase in cui un vecchio schema di attivazione (che non risolve il problema) viene sostituito da un nuovo schema (che lo risolve).

Bowden e Jung-Beeman (2003) hanno mostrato con fMRI e EEG che l'insight è associato a un burst di attivazione nella corteccia temporale destra anteriore circa 300 ms prima del momento consapevole dell'insight, con un'oscillazione gamma preceduta da una riduzione nell'attività alpha (che riflette un "rilassamento" dell'elaborazione focale). Questi pattern sono coerenti con un modello termodinamico in cui:

- La riduzione dell'alpha riflette una riduzione della "temperatura effettiva" del sistema cognitivo — il sistema si stabilizza temporaneamente vicino a uno stato di bassa energia (meno rumore, più cristallizzazione delle strutture esistenti)
- Il burst gamma riflette la transizione rapida verso un nuovo attractor — una biforcazione che porta il sistema verso uno stato di bassa energia libera (soluzione del problema)

La struttura formale è quella di una doppia buca di potenziale nel paesaggio dell'energia libera cognitiva:

$$F(\sigma) = -a/2 \sigma^2 + b/4 \sigma^4$$

dove σ è il "parametro d'ordine" dello stato cognitivo. Per $a < 0$, il sistema ha un unico minimo ($\sigma = 0$, stato di "non-soluzione"). Per $a > 0$, il sistema ha due minimi ($\pm\sqrt{a/b}$, stati di "soluzione"). Il cambiamento di a da negativo a positivo — prodotto da un cambio di prospettiva, dall'esposizione a un hint, o da una riduzione delle inibizioni cognitive — produce la biforcazione verso la soluzione: l'insight.

IX. Conclusioni: la vita come termodinamica della complessità

La termodinamica della vita rivela una struttura fondamentale che percorre l'intero ciclo: la complessità biologica — il cervello, il metabolismo, i cicli biologici, la cognizione — non viola il secondo principio della termodinamica, lo sfrutta. Gli organismi sono strutture dissipative che importano energia libera dall'ambiente, la usano per costruire e mantenere organizzazione interna, e scaricano entropia nell'ambiente come calore.

Il cervello è la struttura dissipativa più complessa prodotta dall'evoluzione: consuma il 20% del metabolismo corporeo per mantenere e operare 10^{14} sinapsi, produce oscillazioni di Turing che codificano e trasmettono informazione, usa l'LTP per costruire modelli generativi dell'ambiente, e richiede il sonno per normalizzare la complessità sinaptica accumulata durante il giorno.

I limiti termodinamici della cognizione biologica non sono i limiti di Landauer (il cervello opera molto lontano da essi) ma i limiti delle leggi di scala metabolica (la dimensione del cervello è vincolata dal budget energetico corporeo), della SHY (la dimensione del cervello è vincolata dalla necessità del sonno per la normalizzazione sinaptica), e della struttura dissipativa (la cognizione richiede flusso continuo di energia — in assenza di glucosio, il cervello perde funzione in minuti).

Comprendere questi limiti fisici — non come ostacoli da superare ma come la struttura entro cui la cognizione biologica si realizza — è parte essenziale di una teoria biologica della mente che sia simultaneamente rigorosa nel senso della fisica e autentica nel senso della biologia.

Monografia XIII

L'evoluzione della cooperazione come matematica degli ecosistemi

Coevoluzione preda-predatore, mutualismi obbligati, la struttura algebrica delle reti ecologiche e i limiti dell'equilibrio spontaneo

I. Oltre la metafora della lotta per la sopravvivenza

La metafora darwiniana della "lotta per la sopravvivenza" ha dominato l'immaginario sulla natura per un secolo e mezzo, producendo interpretazioni che hanno oscillato tra il pessimismo cosmico (la natura è una guerra di tutti contro tutti) e il trionfalismo del mercato (la concorrenza produce il meglio). Entrambe le interpretazioni fraintendono il risultato evolutivo fondamentale: i sistemi biologici più stabili e più produttivi non sono quelli dominati dalla competizione pura, ma quelli in cui la competizione e la cooperazione si bilanciano in strutture ecologiche complesse che la sola selezione individuale non può produrre.

La Monografia XIII del ciclo precedente aveva sviluppato la matematica della cooperazione tra individui — la condizione di Hamilton, l'equazione di Price, i cinque meccanismi di Nowak. Questa monografia affronta il livello superiore: la cooperazione tra specie diverse negli ecosistemi — i mutualismi obbligati, le reti ecologiche, la dinamica preda-predatore come gioco evolutivo, la costruzione della nicchia come forma di cooperazione interspecie. La domanda fondamentale è la stessa del ciclo precedente ma applicata alla scala dell'ecosistema: quali strutture cooperative emergono dall'evoluzione, con quali proprietà formali, e con quali impossibilità strutturali?

II. La dinamica preda-predatore come sistema dinamico

II.1 Il modello di Lotka-Volterra e le sue proprietà

Il modello di Lotka-Volterra (Lotka, 1925; Volterra, 1926) è il punto di partenza di ogni analisi formale delle dinamiche preda-predatore. Le equazioni del modello base per una preda N e un predatore P sono:

$$dN/dt = rN - \alpha NP \quad dP/dt = \beta \alpha NP - \delta P$$

dove r è il tasso di crescita intrinseca della preda, α è il tasso di cattura del predatore (la probabilità di incontro per unità di tempo per individuo), β è l'efficienza di conversione della preda in prole del predatore, e δ è il tasso di mortalità del predatore. Il termine αNP è la funzione di risposta funzionale di tipo I (lineare nella densità della preda).

Il sistema ha due punti fissi:

$P_1 = (0, 0)$: estinzione di entrambe le specie $P_2 = (N^*, P^*) = (\delta/(\beta\alpha), r/\alpha)$: coesistenza

L'analisi di stabilità al punto P_2 usa la matrice jacobiana:

$$J|_{P_2} = \begin{bmatrix} 0 & -\alpha N^* \\ \beta \alpha P^* & 0 \end{bmatrix}$$

con autovalori $\lambda = \pm i\sqrt{r\delta}$ — autovalori puramente immaginari, che indicano che P_2 è un centro neutralmente stabile. Le traiettorie sono orbite chiuse intorno a P_2 — oscillazioni neutrali conservative, come già analizzato nella Monografia XXII del ciclo precedente nel contesto dei giochi coevolutivi.

Questa struttura neutra è un artefatto della linearità del modello. Perturbazioni del modello — termini non lineari realistici — alterano il tipo di stabilità. La risposta funzionale di Holling (1959) generalizza il termine di cattura per includere la saturazione del predatore (un predatore non può mangiare indefinitamente più preda all'aumentare della densità della preda):

Tipo II: $F(N) = \alpha N / (1 + \alpha h N)$ Tipo III: $F(N) = \alpha N^2 / (1 + \alpha h N^2)$

dove h è il "handling time" — il tempo che il predatore impiega per processare ogni preda catturata. Il tipo II produce una risposta lineare a basse densità e saturazione ad alte densità; il tipo III produce un pattern sigmoidale con bassa efficienza a basse densità (effetto di apprendimento) e saturazione ad alte densità.

Con risposta funzionale di tipo II:

$$dN/dt = rN(1 - N/K) - [\alpha N/(1 + \alpha h N)] \cdot P \quad dP/dt = \beta [\alpha N/(1 + \alpha h N)] \cdot P - \delta P$$

dove K è la capacità portante dell'habitat. Il sistema ha un comportamento più ricco: per certi valori dei parametri, il punto di coesistenza diventa un attrattore stabile (spirale stabile); per altri, diventa instabile con un ciclo limite stabile ("Rosenzweig-MacArthur destabilization").

II.2 Il paradosso dell'arricchimento di Rosenzweig

Rosenzweig (1971) dimostrò un risultato controintuitivo: l'arricchimento dell'habitat (aumento della capacità portante K) destabilizza il sistema preda-predatore invece di stabilizzarlo. Aumentando K , il punto di coesistenza P_2 si sposta, l'attività predatoria aumenta, e il sistema può entrare in un regime di oscillazioni di grande ampiezza che conducono all'estinzione stocastica di una o entrambe le specie.

La condizione per la destabilizzazione è:

$$K > K_{\text{critico}} = 1/(\alpha h) + \delta/(\alpha(\beta - \delta h))$$

Per $K > K_{\text{critico}}$, il punto di coesistenza diventa instabile (parte reale degli autovalori positiva) e appare un ciclo limite stabile. Per K molto grande, le oscillazioni hanno ampiezza tale che i minimi di popolazione scendono vicino a zero — e le fluttuazioni stocastiche possono estinguere la specie.

Il paradosso ha una struttura formale che compare anche nella Monografia II di questo ciclo: l'aggiunta di risorse (equivalente all'arricchimento dell'habitat) può destabilizzare un sistema che era in equilibrio stabile. In termini di paesaggio energetico: l'arricchimento aumenta la profondità del minimo di coesistenza ma riduce la sua larghezza, producendo un basin di attrazione più profondo ma più stretto — il sistema è più "determinato" ma meno robusto alle fluttuazioni.

II.3 Il modello a tre specie e il caos deterministico

Hastings e Powell (1991) hanno dimostrato che un sistema di tre specie in catena alimentare lineare (preda $\rightarrow$ consumatore $\rightarrow$ superpredatore) può esibire caos deterministico per certi valori dei parametri. Il sistema:

$$\begin{aligned} dN/dt &= rN(1-N/K) - f_1(N) \cdot C \\ dC/dt &= e_1 f_1(N) \cdot C - f_2(C) \cdot P - d_1 C \\ dP/dt &= e_2 f_2(C) \cdot P - d_2 P \end{aligned}$$

dove f_1 e f_2 sono risposte funzionali di tipo II, produce orbite caotiche (sensibilità alle condizioni iniziali, stretching e folding delle orbite) per parametri ecologicamente realistici.

Il caos deterministico ha conseguenze fondamentali per la previsione delle dinamiche ecologiche: anche in assenza di rumore stocastico, le traiettorie del sistema sono imprevedibili a lungo termine perché piccole perturbazioni si amplificano esponenzialmente. L'esponente di Lyapunov massimale $\lambda_{\max}$ misura questa amplificazione:

$$\|\delta x(t)\| \approx \|\delta x(0)\| \cdot \exp(\lambda_{\max} \cdot t)$$

Per il modello di Hastings-Powell con parametri realistici, $\lambda_{\max} \approx 0.1\text{--}1 \text{ giorno}^{-1}$ — le perturbazioni si amplificano di un fattore e in 1–10 giorni. Questo pone un limite fondamentale alla previsione delle dinamiche di popolazione: anche con modelli perfetti e misure precise, la previsione affidabile è impossibile oltre un orizzonte temporale di $1/\lambda_{\max}$.

III. La struttura algebrica delle reti ecologiche

III.1 Le reti trofiche come grafi orientati

Una rete trofica (food web) è un grafo orientato $G = (V, E)$ dove V è l'insieme delle specie (o dei gruppi funzionali) e E è l'insieme delle relazioni trofiche (archi diretti dalla preda al predatore). Le proprietà topologiche delle reti trofiche hanno struttura sistematica che riflette i vincoli energetici e evolutivi delle comunità ecologiche.

La "scala della comunità" $S = |V|$ (numero di specie), la connettanza $C = |E|/(S^2)$ (frazione di possibili link che sono realizzati), e il link density $L = |E|/S$ (numero medio di link per specie) sono le statistiche di base delle reti trofiche. May (1972) dimostrò che reti casuali con N specie, connettanza C e forza media delle interazioni a (deviazione standard) sono stabili se e soltanto se:

$$a \cdot \sqrt{N \cdot C} < 1$$

Questa condizione — nota come "May's stability criterion" — predice che reti grandi (grande N) o reti molto connesse (grande C) sono instabili. Questo sembra incompatibile con l'osservazione che gli ecosistemi reali, che hanno centinaia di specie e alte connettanze, sono stabili.

La risoluzione del paradosso di May è nella struttura non casuale delle reti reali: le interazioni ecologiche non sono distribuite casualmente ma hanno struttura che favorisce la stabilità. Yodzis (1981) e DeAngelis

(1992) mostrano che reti con struttura ad "intervallo" (le prede di ogni predatore formano un intervallo continuo nello spazio delle nicchie) o con struttura "modulare" (la rete è suddivisibile in moduli poco connessi) sono molto più stabili delle reti casuali con le stesse statistiche di base.

III.2 Il criterio di stabilità di Lyapunov per le reti trofiche

La stabilità di una comunità ecologica all'equilibrio si analizza attraverso la matrice della comunità — la matrice Jacobiana del sistema di equazioni differenziali alle popolazioni di equilibrio:

$$J|_{\{N^*\}} = [\partial(dN_i/dt)/\partial N_j]_{\{N=N^*\}}$$

La comunità è localmente asintoticamente stabile se e soltanto se tutti gli autovalori di J hanno parte reale negativa — se e soltanto se $\text{Re}(\lambda_i) < 0$ per tutti i λ_i eigenvalori di J .

Per reti ecologiche con interazioni di tipo preda-predatore (+/-), competizione (-/-) e mutualismo (+/+), la matrice J ha struttura:

$$J_{ij} = -\sum_k |a_{ik}| \delta_{kj} \text{ (termine diagonale: auto-regolazione)} \quad J_{ij} = \epsilon_{ij} |a_{ij}| \text{ (termine fuori-diagonale: } \epsilon_{ij} = +1 \text{ per } i \text{ beneficiario di } j, -1 \text{ per } i \text{ danneggiato da } j)$$

dove a_{ij} è la forza dell'interazione tra le specie i e j . L'auto-regolazione negativa ($J_{ii} < 0$) favorisce la stabilità; le interazioni mutualistiche ($J_{ij} > 0$ e $J_{ji} > 0$) tendono a destabilizzare.

Il teorema di Gershgorin (algebra lineare) dà un bound sugli autovalori:

$$\text{Re}(\lambda_i) \leq J_{ii} + \sum_{j \neq i} |J_{ij}|$$

La stabilità richiede che i termini diagonali (auto-regolazione) dominino la somma dei termini fuori-diagonale — la stabilità è garantita quando ciascuna specie è sufficientemente auto-regolata rispetto alla forza totale delle sue interazioni con le altre specie.

III.3 La teoria del nesting nelle reti mutualistiche

Le reti mutualistiche pianta-impollinatore e pianta-dispersore di semi mostrano una proprietà strutturale caratteristica: il "nesting" (annidamento). Una rete è nested se le specie con pochi partner (specie specialiste) tendono ad interagire con un sottoinsieme delle specie che interagiscono con le specie con molti partner (le specie generaliste). Formalmente, in una matrice di incidenza binaria A_{ij} (1 se la pianta j

interagisce con l'impollinatore i , 0 altrimenti), il nesting misura quanto il set di partner delle specie meno connesse è contenuto nel set di partner delle specie più connesse.

La misura standard di nesting è il NODF (Nestedness based on Overlap and Decreasing Fill, Almeida-Neto et al. 2008):

$$\text{NODF} = [\sum_{i < j} (MT_{ij} + PO_{ij})] / [C(n_r, 2) + C(n_c, 2)]$$

dove MT_{ij} è 1 se la riga i ha più interazioni della riga j (marginal total), PO_{ij} è la percentuale di overlap nelle connessioni tra riga i e riga j , n_r e n_c sono il numero di righe e colonne, e $C(\cdot, \cdot)$ sono i coefficienti binomiali.

Bascompte et al. (2003) hanno mostrato che le reti mutualistiche reali hanno NODF significativamente più alto di reti casuali con la stessa dimensione e connettanza. Questa struttura nested non è un artefatto ma una proprietà adattiva: le reti nested sono più robuste alla perdita di specie (l'estinzione di una specie specialista non causa effetti a cascata, mentre le specie generaliste sono le ultime a estinguersi e la loro perdita colpisce molte specie specialiste che dipendono da loro).

La struttura nested ha anche un'interpretazione evoluzionista: emerge naturalmente dal processo di colonizzazione e coevoluzione dove le nuove specie che colonizzano un sistema tendono ad interagire con le specie più abbondanti (effetto di frequenza) producendo un pattern di inclusione degli specialisti nei generalisti.

IV. I mutualismi obbligati e la coevoluzione

IV.1 La struttura formale del mutualismo obbligato

Un mutualismo obbligato è una relazione in cui due specie dipendono strettamente l'una dall'altra per sopravvivere e riprodursi — nessuna delle due può persistere senza l'altra. Gli esempi classici includono: la simbiosi degli zooxantelle (alghe fotosintetiche) con i coralli (che dipendono dalle alghe per il 90% del loro fabbisogno calorico); la simbiosi dei rizobi (batteri azoto-fissatori) con le leguminose; la simbiosi dell'*Yucca* con la tarma *Yucca* (l'unica pollinatrice della *Yucca*, che depone le uova nei fiori che impollina); e la simbiosi delle formiche *Atta* con i funghi che coltivano nei loro nidi.

Il modello matematico del mutualismo obbligato è una generalizzazione del modello di Lotka-Volterra con termini di beneficio reciproco:

$$dN_1/dt = N_1(r_1 - d_1N_1 + \alpha_{12}N_2) \quad dN_2/dt = N_2(r_2 - d_2N_2 + \alpha_{21}N_1)$$

dove r_1 e r_2 sono tassi di crescita intrinseca (che possono essere negativi per mutualismi obbligati: senza il partner la specie declina), d_1 e d_2 sono coefficienti di auto-limitazione, e $\alpha_{12}, \alpha_{21} > 0$ sono i tassi di beneficio mutualistico.

Per mutualismi obbligati ($r_1 < 0$ e $r_2 < 0$ — entrambe le specie hanno crescita negativa in assenza del partner), il sistema ha tre possibili comportamenti a seconda dei parametri:

- (1) Estinzione di entrambe: se $\alpha_{12}\alpha_{21} < d_1d_2$ (il beneficio mutualistico non è sufficiente a compensare la mortalità intrinseca), l'unico punto fisso stabile è l'origine.
- (2) Coesistenza stabile: se $\alpha_{12}\alpha_{21} > d_1d_2$ e $r_1r_2 > 0$ (stesso segno), esiste un punto fisso di coesistenza stabile con entrambe le specie presenti.
- (3) Crescita illimitata ("mutualismo runaway"): per certi valori dei parametri, il beneficio mutualistico può produrre crescita esplosiva — il sistema è instabile e le popolazioni crescono verso l'infinito.

La condizione di coesistenza stabile per mutualismi obbligati è più restrittiva di quella per la competizione: richiede che il beneficio mutualistico sia sufficientemente forte da compensare la mortalità intrinseca, ma non così forte da produrre la crescita illimitata. Questa "sweet spot" della cooperazione obbligata è una struttura fragile — facilmente perturbata da cambiamenti ambientali che alterano l'efficacia del beneficio mutualistico (come il bleaching dei coralli che perde le zooxantelle quando la temperatura aumenta).

IV.2 La coevoluzione dei mutualismi e i "fitment landscapes"

I mutualismi coevolvono: le specie partecipanti si adattano reciprocamente attraverso la selezione, producendo una co-speciation (speciation parallela delle due specie mutualistiche) o una diversificazione differenziale. Thompson (1994, 2005) ha proposto la "geographic mosaic theory of coevolution": la coevoluzione non avviene uniformemente nella distribuzione geografica delle specie interagenti, ma è eterogenea — certe regioni sono "hotspot" coevolutivi (dove la selezione reciproca è forte), altre sono "coldspot" (dove la selezione reciproca è debole o assente).

La formalizzazione usa la struttura dei "trait landscapes" coevolutivi. Siano x_1 e x_2 i traits dell'host e del symbionte (ad esempio la dimensione del fiore e la lunghezza della tromba dell'impollinatore nella coppia fiore-impollinatore). Il "matching fitness" — la fitness di entrambe le specie in funzione dei traits — è:

$W_1(x_1, x_2) = f_1(x_1 - x_2)$ (fitness dell'host: massima quando il trait del symbionte "fitta" il proprio) $W_2(x_1, x_2) = f_2(x_2 - x_1)$ (fitness del symbionte: analoga)

dove f_1 e f_2 sono funzioni decrescenti con la distanza $|x_1 - x_2|$ (la fitness è massima quando i traits si corrispondono). La coevoluzione è la dinamica nel piano (x_1, x_2) guidata dai gradienti della fitness:

$$dx_1/dt = h_1 \cdot G_1 \cdot \partial W_1 / \partial x_1 = -h_1 G_1 \cdot f_1'(x_1 - x_2) \quad dx_2/dt = h_2 \cdot G_2 \cdot \partial W_2 / \partial x_2 = h_2 G_2 \cdot f_2'(x_2 - x_1)$$

dove h_1, h_2 sono la ereditabilità dei traits e G_1, G_2 sono le varianze genetiche additive. Il sistema converge a (x_1^*, x_2^*) con $x_1^* = x_2^*$ — il matching perfetto — ma la traiettoria di convergenza dipende dai valori relativi di $h_1 G_1$ e $h_2 G_2$. Se una specie evolve molto più rapidamente dell'altra, la traiettoria è asimmetrica e la convergenza può essere molto lenta.

V. Le reti di mutualismo come strutture algebriche

V.1 Il problema della stabilità dei mutualismi in comunità

Un risultato sorprendente delle analisi teoriche dei mutualismi nelle comunità è che le interazioni mutualistiche tendono a destabilizzare le comunità — aumentando i valori massimi degli autovalori della matrice della comunità J. May (1972) e analisi successive mostrano che, a parità di connettanza e intensità media delle interazioni, comunità con più mutualismi sono meno stabili di comunità con più competizioni.

Il paradosso è che i mutualismi sono ubiquitari negli ecosistemi reali (si stima che più del 50% delle interazioni nella maggior parte degli ecosistemi terrestri siano mutualistiche — impollinazione, dispersione di semi, micorrize) e che questi ecosistemi siano tuttavia stabili. La risoluzione del paradosso, proposta da Holland e DeAngelis (2010) e Rohr et al. (2014), è che la struttura nested dei mutualismi — non i mutualismi di per sé — produce stabilità: le reti mutualistiche nested

hanno un insieme di autovalori dominanti della matrice J che è più negativo della media prevista per reti casuali con le stesse statistiche di base.

Rohr et al. (2014) dimostrano formalmente che la struttura nested di una rete mutualistica produce un pattern specifico nella matrice J : le righe e colonne corrispondenti alle specie generaliste (con molte interazioni) hanno elementi negativi grandi (forte auto-regolazione effettiva), mentre le righe e colonne delle specie specialiste hanno elementi negativi piccoli. La dominanza degli autovalori negativi derivanti dalle specie generaliste stabilizza l'intero sistema.

La formalizzazione usa la decomposizione spettrale della matrice J della comunità:

$$J = Q \Lambda Q^{-1}$$

dove Λ è la matrice diagonale degli autovalori e Q è la matrice degli autovettori. Il requisito di stabilità $\text{Re}(\lambda_i) < 0 \quad \forall i$ è soddisfatto se la struttura nested della rete mutualistic produce una predominanza di autovalori con parte reale fortemente negativa — e Rohr et al. mostrano analiticamente che questo è il caso per strutture nested con abbastanza generalisti nel sistema.

V.2 Il teorema della "bio-diversity begets stability"

Il dibattito sulla relazione tra diversità biologica e stabilità degli ecosistemi è uno dei più longevi nell'ecologia teorica. Elton (1958) propose intuitivamente che gli ecosistemi più diversi sono più stabili; May (1972) dimostrò formalmente che reti casuali diventano meno stabili all'aumentare della diversità. La risoluzione di questo apparente paradosso viene da McCann, Hastings e Huxel (1998) e Tilman et al. (2006), che mostrano che la relazione diversità-stabilità dipende crucialmente dal tipo di diversità e dal tipo di stabilità considerato.

Per la stabilità di resistenza (la capacità di mantenere il livello di biomassa dopo una perturbazione), Tilman et al. (2006) mostrano empiricamente e teoricamente che la diversità aumenta la stabilità per le comunità di produttori primari (piante), attraverso il meccanismo della "portafoglio di specie" (portfolio effect): con specie diverse che rispondono in modo diverso alle perturbazioni, le fluttuazioni si compensano parzialmente, riducendo la varianza totale della biomassa.

Il "portfolio effect" è la versione ecologica del teorema di Markowitz per la diversificazione del rischio finanziario. Sia σ^2_{pop} la varianza della biomassa totale di una comunità con N specie con varianze individuali σ^2_i e correlazioni ρ_{ij} tra le fluttuazioni:

$$\sigma^2_{\text{pop}} = \sum_i \sigma^2_i / N^2 + 2 \sum_{\{i < j\}} \rho_{\{ij\}} \sigma_i \sigma_j / N^2$$

Per correlazioni basse ($\rho_{\{ij\}} \approx 0$) e varianze uguali ($\sigma_i = \sigma$):

$$\sigma^2_{\text{pop}} \approx \sigma^2 / N$$

La varianza della biomassa totale decresce come $1/N$ — la stabilità aumenta con la diversità, e la radice della riduzione è proprio il portfolio effect. Per $\rho_{\{ij\}} > 0$ (specie correlate nelle risposte), il portfolio effect è ridotto; per $\rho_{\{ij\}} < 0$ (specie anticorrelate, che rispondono in modo opposto), il portfolio effect è amplificato.

VI. L'ecologia del suolo come sistema cognitivo distribuito

VI.1 La rete miceliare come "internet del bosco"

Le reti di micorrize — associazioni mutualistiche tra i funghi (micorrize) e le radici delle piante — formano reti fisiche che connettono piante diverse nello stesso suolo, permettendo il trasferimento di carbonio, azoto, fosforo e segnali chimici tra piante a distanza. Simard et al. (1997) hanno dimostrato sperimentalmente che il carbonio fissato dalla fotosintesi di un albero viene trasferito attraverso la rete miceliare a alberi vicini — inclusi alberi di specie diverse — con tassi che dipendono dalla struttura topologica della rete.

Questa struttura — battezzata "wood wide web" da Helgason et al. (1998) — ha la struttura algebrica di un grafo pesato $G_{\text{funghi}} = (V_{\text{piante}}, E_{\text{ife}}, W_{\text{transfert}})$ dove V_{piante} è l'insieme delle piante connesse, E_{ife} sono le connessioni fungine tra piante, e $W_{\text{transfert}}$ è la matrice dei tassi di trasferimento di carbonio. Il grafo ha proprietà small-world: pochi "hub" fungini con molte connessioni connettono molte piante attraverso percorsi brevi.

La formalizzazione del trasferimento di carbonio usa il modello di diffusione sulla rete:

$$dC_i/dt = F_i - \sum_j T_{ij} (C_i - C_j)$$

dove C_i è la concentrazione di carbonio nella pianta i , F_i è il tasso di fissazione fotosintetica, e T_{ij} è il tasso di trasferimento attraverso la rete miceliare tra i e j . Il sistema tende verso uno stato stazionario in cui il carbonio è distribuito tra le piante proporzionalmente alle loro

necessità metaboliche — le piante più in ombra (con bassa fotosintesi) ricevono carbonio dalle piante in luce piena (con alta fotosintesi).

Questo trasferimento non è passivo (semplice diffusione da alta a bassa concentrazione) ma è regolato attivamente dal fungo: i funghi micorrizici favoriscono le piante che forniscono più zuccheri, creando una struttura di preferenza commerciale che massimizza il flusso di carbonio verso il fungo. La struttura "di mercato" che emerge — con il fungo come intermediario e il prezzo del trasferimento determinato dalla domanda-offerta di carbonio e nutrienti — è un esempio di cooperazione strutturata che emerge da un'interazione apparentemente puramente selfish.

VI.2 Il microbioma come ecosistema cognitivo

Il microbioma intestinale umano — circa 10^{13} – 10^{14} batteri, virus, funghi e altri microrganismi nell'intestino — forma un ecosistema interno che ha effetti documentati sulla cognizione attraverso l'asse intestino-cervello (gut-brain axis). La comunicazione tra microbioma e cervello avviene attraverso il nervo vago (trasmissione nervosa diretta), attraverso ormoni enterici (serotonina, dopamina, GABA prodotti dall'intestino), attraverso la risposta immunitaria (citochine che attraversano la barriera emato-encefalica), e attraverso la produzione di metaboliti che influenzano la funzione neurale.

Cryan e Dinan (2012) hanno documentato che alterazioni del microbioma (antibiotici, dieta, stress) producono cambiamenti nel comportamento in modelli animali — aumento dell'ansia, riduzione della socievolezza, cambiamenti nelle risposte allo stress — e che il trapianto del microbioma da topi "audaci" a topi "timidi" produce parziale trasmissione del fenotipo comportamentale. Il microbioma non è soltanto un insieme di batteri che digerono il cibo: è un sistema ecologico che co-determina gli stati del cervello attraverso canali di comunicazione multipli.

La struttura ecologica del microbioma è descrivibile con gli stessi strumenti delle reti ecologiche: il microbioma è una comunità con struttura di rete trofica (alcune specie batteriche mangiano i prodotti metabolici di altre), con mutualismi (batteri che co-dipendono per la sintesi di nutrienti essenziali), con competizione (specie che competono per gli stessi substrati), e con keystone species (specie le cui estinzioni producono collasso della comunità).

VII. L'emergenza della cooperazione come problema di teoria dei giochi evolutivi

VII.1 La struttura degli ecosistemi come gioco di coordinazione

Le interazioni ecologiche tra specie diverse possono essere formalizzate come giochi evolutivi in cui le strategie (le nicchie ecologiche, i tratti coevolutivi) vengono selezionate in base al loro payoff relativo nella comunità. La novità rispetto ai giochi tra individui della stessa specie (analizzati nella Monografia XIII del ciclo precedente) è che i payoff dipendono dalle frequenze non soltanto della stessa specie ma di specie diverse — la dinamica è multi-specie.

Per un sistema con due specie che interagiscono attraverso una risorsa condivisa, la struttura del gioco è quella di un commons game (gioco dei beni comuni):

Specie A: estrattore $\rightarrow$ payoff $\alpha N_A - \beta N_B$ Specie B: estrattore $\rightarrow$ payoff $\alpha N_B - \gamma N_A$

dove il termine βN_B rappresenta la competizione interspecie (A è danneggiata dalla presenza di B che sfrutta la stessa risorsa) e γN_A rappresenta l'interazione dalla parte di B. Se $\beta > 0$ e $\gamma > 0$ (competizione), il gioco è di tipo "pollo" con payoff simmetrico. Se $\beta > 0$ e $\gamma < 0$ (A compete con B, ma B trae beneficio da A — una relazione asimmetrica di tipo commensalismo o amensalismo), la struttura del gioco cambia.

La generalizzazione a sistemi con n specie produce la matrice di payoff:

$$W_i(N_1, \dots, N_n) = r_i + \sum_j \alpha_{ij} N_j$$

dove α_{ij} è il coefficiente di interazione (positivo per mutualismo, negativo per competizione, zero per neutralità). La dinamica evolutiva è governata dall'equazione di replicazione generalizzata:

$$dN_i/dt = N_i \cdot (W_i(N) - \bar{W}(N))$$

dove $\bar{W} = \sum_i N_i W_i / \sum_i N_i$ è il payoff medio. Gli equilibri di questa dinamica sono i punti fissi del sistema — gli stati in cui nessuna specie può aumentare la propria abbondanza.

VII.2 L'impossibilità dell'ottimo globale negli ecosistemi

Il teorema che chiude questa monografia parallela il paradosso di Rogers (Monografia VII) e l'impossibilità del Nash nel MARL (Monografia XXII del ciclo precedente): in ecosistemi con interazioni competitive asimmetriche, la dinamica evolutiva non converge all'ottimo collettivo (la configurazione che massimizza la produzione totale dell'ecosistema), ma a un insieme di equilibri di Nash che possono essere molto sub-ottimali rispetto all'ottimo collettivo.

La "paradox of enrichment" di Rosenzweig è un caso speciale di questo impossibilità: l'aggiunta di risorse all'ecosistema (che dovrebbe aumentare la produzione totale) invece destabilizza la struttura e può portare all'estinzione. La struttura della ricerca del proprio interesse individuale (massimizzazione dell'idoneità di ciascuna specie) non produce necessariamente l'ottimo collettivo (la struttura dell'ecosistema più produttiva).

Questo risultato non è sorprendente alla luce dei teoremi di impossibilità della teoria dei giochi. Ma ha una conseguenza importante: la diversità biologica osservata negli ecosistemi reali non è necessariamente "ottimale" — è il risultato di processi evolutivi che hanno prodotto equilibri locali, dipendenti dalla storia evolutiva (path dependence) e spesso sub-ottimali rispetto all'ottimo della produzione collettiva. La biodiversità che vediamo è la soluzione raggiunta da un algoritmo di ottimizzazione senza garante del ottimo globale, operante per miliardi di anni in un paesaggio adattativo che cambia continuamente.

VIII. Conclusioni: l'ecosistema come struttura di cooperazione necessaria

Il percorso di questa monografia ha mostrato come la matematica degli ecosistemi — le reti trofiche, i mutualismi obbligati, le strutture nested, le reti micorrize — non sia semplicemente la descrizione quantitativa di fenomeni biologici, ma riveli strutture algebriche e topologiche profonde che determinano le proprietà di stabilità, diversità e produttività degli ecosistemi.

La cooperazione negli ecosistemi non emerge perché le specie "decidono" di cooperare — emerge perché le strutture coevolutive che producono mutualismi, commensalismi e costruzione della nicchia sono

le strutture che sopravvivono alla selezione nell'ambiente competitivo. L'ecosistema è una struttura di cooperazione necessaria — non scelta, ma imposta dalla termodinamica e dalla struttura evolutiva dei paesaggi adattativi.

I limiti di questa cooperazione sono i limiti che la matematica rivela: la soglia di Rosenzweig (l'arricchimento destabilizza), la fragilità dei mutualismi obbligati (la sweet spot è stretta), il caos deterministico nelle catene alimentari lunghe (la previsione è impossibile oltre l'orizzonte di Lyapunov). Questi non sono difetti degli ecosistemi — sono le proprietà necessarie di sistemi che operano al margine della stabilità, sufficientemente vicini all'instabilità da essere flessibili e adattativi, ma non così lontani da collassare per le perturbazioni ordinarie. Comprendere questo margine è comprendere la logica profonda della vita come fenomeno fisico e biologico.

Monografia XIV

La neurobiologia del valore e della decisione

Il sistema dopaminergico come reward prediction error, la neuroscienza della scelta, i limiti computazionali del cervello decisionale

I. Il cervello che sceglie: dall'aristotelismo alla neuroscienza

Aristotele descrisse la deliberazione pratica — il processo attraverso cui l'agente razionale sceglie le proprie azioni — come un sillogismo pratico: la premessa maggiore è il desiderio (boulesis, la volontà di raggiungere un certo fine), la premessa minore è la credenza (le condizioni che permettono di raggiungere quel fine), e la conclusione è l'azione. Questo schema — semplice nella forma, profondo nelle implicazioni — anticipa di duemilaquattrocento anni la struttura che la neuroscienza moderna ha scoperto nel cervello: un sistema che valuta le opzioni disponibili rispetto a obiettivi impliciti, aggiorna quella valutazione sulla base dell'esperienza, e seleziona l'azione con il valore atteso più alto.

Ma la neuroscienza ha scoperto qualcosa che Aristotele non poteva anticipare: il sistema di valutazione del cervello non è un singolo processo razionale unificato — è un insieme di sistemi parzialmente indipendenti che competono e si integrano, ciascuno con la propria architettura computazionale, ciascuno con i propri limiti strutturali, e ciascuno con le proprie vulnerabilità patologiche. La "razionalità" del cervello decisionale non è la razionalità dell'ottimizzatore di von Neumann-Morgenstern: è la razionalità ecologica — l'insieme di soluzioni che l'evoluzione ha selezionato per funzionare bene negli ambienti in cui la specie si è evoluta, con le inevitabili distorsioni quando quegli ambienti cambiano.

II. Il sistema dopaminergico e il reward prediction error

II.1 La scoperta di Schultz

Wolfram Schultz (1997, 1998) ha prodotto uno dei risultati più citati nella neuroscienza degli ultimi trent'anni: i neuroni dopaminergici del mesencefalo (area ventrale tegmentale, VTA, e substantia nigra pars compacta, SNc) non codificano la ricompensa assoluta come si credeva, ma codificano l'errore di previsione della ricompensa (reward prediction error, RPE) — la differenza tra la ricompensa ottenuta e la ricompensa attesa.

L'evidenza fondamentale viene da esperimenti di condizionamento Pavloviano in primati. Un tono (CS, conditioned stimulus) viene presentato seguito da succo di frutto (US, unconditioned stimulus). Prima del condizionamento, i neuroni DA scaricano in risposta al succo (la ricompensa inattesa). Dopo il condizionamento:

Prima del tono: scarica basale Durante il tono: scarica aumentata (il tono predice la ricompensa) Al tempo del succo: scarica invariata (la ricompensa è attesa — nessun errore) Se il succo è omesso: scarica ridotta sotto basale (errore negativo: la ricompensa attesa non è arrivata)

Il pattern di risposta è:

$$\delta(t) = r(t) - V(t|S_t)$$

dove $r(t)$ è la ricompensa ricevuta al tempo t , $V(t|S_t)$ è la ricompensa attesa sulla base dello stimolo corrente S_t , e $\delta(t)$ è l'RPE. Il sistema dopaminergico codifica $\delta(t)$ con la sua scarica:

$\delta > 0 \rightarrow$ aumento della scarica sopra la basale (ricompensa migliore del previsto) $\delta = 0 \rightarrow$ scarica basale invariata (ricompensa come previsto) $\delta < 0 \rightarrow$ riduzione della scarica sotto la basale (ricompensa peggiore del previsto)

II.2 La connessione con il TD learning

L'RPE di Schultz corrisponde esattamente all'errore di TD (Temporal Difference) nell'algoritmo di apprendimento per rinforzo TD(λ) di Sutton (1988). L'algoritmo TD(0) aggiorna la stima $V(s)$ del valore dello stato s come:

$$V(s_t) \leftarrow V(s_t) + \eta \cdot \delta_t$$

dove:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

è l'errore TD. Il termine r_{t+1} è la ricompensa al passo successivo, γ è il fattore di sconto per le ricompense future ($0 < \gamma < 1$), $V(s_{t+1})$ è il valore stimato dello stato successivo, e $V(s_t)$ è il valore stimato dello stato corrente.

La corrispondenza tra δ_t (TD) e il segnale dopaminergico $\delta(t)$ (Schultz) è:

$$r_{t+1} \leftrightarrow r(t) \text{ (ricompensa immediata)} \quad \gamma V(s_{t+1}) \leftrightarrow V_{\text{futuro}}(t) \text{ (valore futuro stimato)} \\ V(s_t) \leftrightarrow V(t|S_t) \text{ (valore correntemente previsto)}$$

Il segnale dopaminergico implementa il TD(0) con una precisione che non ha precedenti nella neuroscienza computazionale: è il primo caso in cui un algoritmo di apprendimento per rinforzo sviluppato indipendentemente nell'informatica è stato trovato implementato fedelmente in un sistema biologico.

II.3 La distribuzione anatomica della dopamina e la struttura del sistema di valore

Il sistema dopaminergico non è un sistema uniforme: i neuroni DA del VTA e della SNc proiettano a strutture diverse con funzioni diverse, formando due pathway principali:

La via mesolimbica (VTA $\rightarrow$ nucleo accumbens, amigdala, ippocampo, corteccia prefrontale mediale) è la via del "reward learning" — l'apprendimento delle associazioni stimolo-ricompensa e la motivazione a perseguire le ricompense apprese.

La via nigrostriatale (SNc → striato dorsale, putamen, caudato) è la via del "habit learning" — l'apprendimento delle sequenze motorie abituali e la selezione automatica delle azioni.

Questa doppia struttura corrisponde alla distinzione computazionale tra apprendimento goal-directed (orientato agli obiettivi: l'agente valuta le azioni sulla base dei loro esiti attesi) e apprendimento habitual (basato sulle abitudini: l'agente seleziona le azioni in base alle associazioni stimolo-risposta apprese). Daw, Niv e Dayan (2005) hanno proposto che il cervello implementi due sistemi separati di apprendimento per rinforzo:

Sistema model-based (via mesolimbica + corteccia prefrontale): mantiene un modello del mondo (le transizioni tra stati e i valori degli stati), usa questo modello per pianificare le azioni attraverso la ricerca dell'albero decisionale, è flessibile ma computazionalmente costoso.

Sistema model-free (via nigrostriatale): impara direttamente le associazioni stimolo-azione attraverso l'RPE, è rigido (non si aggiorna al cambiamento del valore dell'esito) ma computazionalmente economico.

III. L'architettura neurale della scelta

III.1 Il modello dell'accumulatore e la struttura della decisione percettiva

Il processo di decisione percettiva — come decidere se uno stimolo ambiguo appartiene alla categoria A o alla categoria B — è descritto dal modello dell'accumulatore (drift-diffusion model, DDM; Ratcliff, 1978). In questo modello, l'agente accumula evidenza per le due alternative nel tempo, e sceglie l'alternativa per cui la soglia di decisione viene raggiunta per prima.

Il DDM per la decisione binaria è un processo stocastico:

$$dx = \mu dt + \sigma dW$$

dove $x(t)$ è il livello di evidenza accumulata al tempo t (iniziando da $x(0) = z$, il bias iniziale), μ è il tasso di deriva (l'evidenza media per unità di tempo, proporzionale alla qualità del segnale), σ^2 è il coefficiente di diffusione (il rumore), e dW è il rumore bianco (incremento del processo di Wiener). La decisione viene presa quando $x(t)$ raggiunge la soglia superiore $+a$ (risposta A) o la soglia inferiore $-b$ (risposta B).

Le previsioni del DDM includono: La distribuzione dei tempi di risposta per risposte corrette ed errate La curva psicomedica (proporzione di risposte corrette in funzione del segnale μ) L'effetto di velocità-accuratezza (speed-accuracy trade-off): alzare le soglie ($\pm a, \pm b$) aumenta la proporzione di risposte corrette ma aumenta anche il tempo di risposta.

Il substrato neurale del DDM è identificato nelle aree della corteccia parietale posteriore (area LIP nelle scimmie) e della corteccia prefrontale: i neuroni di LIP mostrano attività che cresce nel tempo durante la decisione percettiva, con la velocità di crescita proporzionale alla qualità del segnale (il μ del DDM) e con decisione quando l'attività raggiunge una soglia (la a del DDM).

III.2 La corteccia orbitofrontale e la codifica del valore

La corteccia orbitofrontale (OFC) è l'area corticale più direttamente associata alla codifica del valore dei stimoli — non il valore primario biologico (il glucosio, la sicurezza, il sesso), ma il valore appreso attraverso l'esperienza e il contesto. I neuroni dell'OFC mostrano attività modulata dal valore degli stimoli in modo separato dalla risposta motoria: lo stesso neurone risponde allo stesso stimolo con intensità diversa a seconda del valore corrente dello stimolo (che dipende dalla storia di rinforzo recente, dal livello di sazietà, dal contesto sociale).

Rolls (2000) e Padoa-Schioppa e Assad (2006) hanno mostrato che l'OFC mantiene rappresentazioni "cardinali" dei valori (non ordinal, ma numeriche) che permettono il confronto tra alternative di diversa categoria. Un neurone dell'OFC che risponde a "un succo di frutto" con una certa intensità risponde a "due succhi di frutto" con il doppio dell'intensità — la codifica è numerica, non semplicemente categoriale.

La formalizzazione usa il concetto di "utility neurons" — neuroni che codificano l'utilità soggettiva di uno stimolo (non il suo valore biologico bruto). La funzione di risposta è:

$$r_{\text{neuron}} = f(U(\text{stimolo}, \text{contesto}))$$

dove U è l'utilità soggettiva — una funzione del valore biologico del stimolo e del contesto (livello di sazietà, rango nella gerarchia sociale, storia recente di ricompense). La forma di f è tipicamente logaritmica ($r \propto \log(U)$) in accordo con la legge di Weber-Fechner della psicofisica.

III.3 Il sistema accumbens-amigdala e la struttura motivazionale

Il nucleo accumbens (NAcc) e l'amigdala basolaterale (BLA) formano un circuito che integra la valutazione degli stimoli (BLA) con la motivazione ad agire per ottenerli o evitarli (NAcc). Il circuito BLA-NAcc è il substrato neurale della "motivazione incentivale" — la forza che spinge verso le ricompense e lontano dalle minacce.

Berridge e Robinson (1998) hanno proposto la distinzione cruciale tra "wanting" (volere) e "liking" (piacere): il sistema dopaminergico-NAcc media il wanting — la motivazione a perseguire una ricompensa — mentre un sistema oppioide separato media il liking — il piacere soggettivo di ricevere la ricompensa. Le due componenti sono dissociabili: topi con distruzione selettiva del sistema dopaminergico non mangiano spontaneamente (ridotto wanting) ma mostrano normali reazioni facciali di piacere quando il cibo viene messo in bocca (normale liking). Topi con somministrazione di oppioidi nel NAcc mediale mostrano aumento del liking (più reazioni facciali positive) senza aumento del wanting (stessa quantità di cibo consumata spontaneamente).

La separazione wanting/liking è fondamentale per comprendere le dipendenze: nella dipendenza da droghe, il sistema dopaminergico viene "dirottato" dalle droghe che producono massicce scariche di dopamina, producendo wanting estremo per le droghe. Ma il liking dalla droga — il piacere soggettivo — si riduce progressivamente con l'uso continuato (tolleranza), mentre il wanting rimane elevato o aumenta (sensitizzazione). Questo produce la struttura paradossale dell'addizione: il desiderio intenso della droga (wanting massimo) in assenza di piacere al suo consumo (liking ridotto).

IV. Il modello bayesiano della decisione e i suoi limiti biologici

IV.1 L'ottimizzazione bayesiana della decisione

Nel framework bayesiano ottimale per la decisione (Bergen e Pouget, 2012; Ma et al., 2006), l'agente mantiene una distribuzione di probabilità sullo stato del mondo e sceglie l'azione che massimizza l'utilità attesa:

$$a^* = \operatorname{argmax}_a \int U(a, s) \cdot p(s | x) ds$$

dove $U(a, s)$ è l'utilità di prendere l'azione a nello stato del mondo s , e $p(s | x)$ è la distribuzione posteriore sullo stato del mondo data l'osservazione x .

Per il caso della decisione binaria (stimolo A o B), questo si riduce al test likelihood ratio:

$$LR(x) = p(x | s=A) / p(x | s=B) \geq \lambda \rightarrow \text{scegli A}$$

dove $\lambda = [U(B,B) - U(A,B)] / [U(A,A) - U(B,A)]$ è il threshold decisionale ottimale. La soglia λ dipende dai costi relativi degli errori di tipo I (scegliere B quando A è vero) e di tipo II (scegliere A quando B è vero).

Il DDM implementa esattamente questo test likelihood ratio in modo sequenziale: l'accumulo dell'evidenza $x(t)$ è equivalente al calcolo del log-likelihood ratio cumulativo:

$$x(t) = \sum_{\tau=0}^t \log LR(y_{\tau})$$

dove y_{τ} è l'evidenza al passo τ . La soglia di decisione λ del DDM corrisponde alla soglia λ del test Bayesiano ottimale — il DDM è l'implementazione sequenziale ottimale del test likelihood ratio in presenza di segnale con rumore gaussiano (Sequential Probability Ratio Test, SPRT, Wald 1947).

IV.2 Il problema del credito temporale e i limiti della pianificazione

Il sistema di apprendimento per rinforzo del cervello — il TD learning implementato dal sistema dopaminergico — ha un limite strutturale fondamentale: il "problema del credito temporale" (credit assignment problem). Quando l'azione di un agente produce una ricompensa molto tempo dopo l'azione, il sistema di TD deve distribuire il "credito" della ricompensa all'indietro nel tempo attraverso le azioni che hanno contribuito a essa.

Il TD(λ) di Sutton (1988) risolve questo problema attraverso l'eligibility trace — una traccia di attività che "ricorda" quali stati e azioni sono stati visitati di recente. Il TD(λ) con eligibility trace z_t è:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad z_t = \gamma \lambda z_{t-1} + \nabla V(s_t) \quad \theta \leftarrow \theta + \eta \delta_t z_t$$

dove z_t è il vettore eligibility trace, $\lambda \in [0,1]$ è il parametro di trace decay, e θ sono i parametri del modello di valore. Per $\lambda = 0$, il TD(0) aggiorna soltanto lo stato corrente; per $\lambda = 1$, il TD(1) è equivalente al Monte Carlo return — distribuisce il credito uniformemente su tutti gli stati visitati.

Il limite biologico è che il sistema dopaminergico implementa qualcosa di più simile al TD(0) o al TD(λ) con λ piccolo: la scarica dei neuroni DA risponde agli RPE in modo locale nel tempo, con scarsa propagazione del credito agli stati distanti nella catena causale. Questo produce la dipendenza all'orizzonte temporale — gli agenti (umani e animali) attribuiscono poco valore alle ricompense temporalmente distanti, con un tasso di sconto che tipicamente segue una curva iperbolica (non esponenziale come la teoria della decisione normativa richiederebbe):

$$V(\text{ricompensa ritardata}) = V(\text{ricompensa immediata}) / (1 + k \cdot t)$$

dove k è il tasso di sconto iperbolico e t è il ritardo. Il discounting iperbolico viola la consistenza temporale — l'agente preferisce la ricompensa piccola-immediata rispetto alla ricompensa grande-ritardata, ma inverte la preferenza quando entrambe le ricompense sono poste nel futuro (preferring-larger-later when both are delayed). Questa inversione della preferenza produce la struttura della procrastinazione e dell'impulsività.

V. La neuroscienza della scelta sociale

V.1 La fairness e il sistema di punishment

Il cervello umano non ottimizza soltanto il proprio payoff — è sensibile alla distribuzione dei payoff tra sé e gli altri. L'Ultimatum Game (UG) è il paradigma sperimentale classico: un proponente riceve una somma e propone come dividerla con un rispondente; il rispondente può accettare (entrambi ricevono la propria quota) o rifiutare (entrambi ricevono zero). La teoria classica predice che il rispondente razionale dovrebbe accettare qualsiasi offerta positiva — anche 1 euro su 10 è meglio di zero. Ma empiricamente, i rispondenti rifiutano sistematicamente offerte percepite come ingiuste (tipicamente offerte < 30% del totale), anche a costo del proprio guadagno.

Sanfey et al. (2003) hanno mostrato con fMRI che le offerte ingiuste attivano l'insula anteriore (correlata con il disgusto e le emozioni negative) e la corteccia prefrontale dorsolaterale (correlata con il controllo cognitivo). Il conflitto tra la risposta emotiva (rifiutare l'offerta ingiusta per indignazione) e la risposta razionale (accettare per massimizzare il guadagno) si manifesta come competizione di attivazione tra insula e DLPFC: i rispondenti che rifiutano le offerte ingiuste mostrano maggiore attivazione dell'insula; quelli che accettano mostrano maggiore attivazione della DLPFC.

La struttura neurobiologica corrisponde al framework dei processi duali di Kahneman (Monografia VI): il Sistema 1 (insula, risposta emotiva immediata) produce la risposta di rifiuto per disgusto morale; il Sistema 2 (DLPFC, ragionamento esplicito) può sovrascrivere questa risposta se il costo dell'accettazione è valutato come superiore al beneficio del rifiuto.

V.2 La teoria dell'equità come funzione di utilità social

Fehr e Schmidt (1999) hanno proposto un modello formale dell'inequity aversion che cattura il comportamento osservato nell'UG e in altri giochi sociali. Il modello propone che l'utilità di un agente i in un gioco con n giocatori sia:

$$U_i(x) = x_i - \alpha_i \cdot (1/(n-1)) \cdot \sum \max(x - x_i, 0) - \beta_i \cdot (1/(n-1)) \cdot \sum \max(x_i - x, 0)$$

dove x_i è il payoff materiale di i , il secondo termine è la "disutilità dell'essere svantaggiato" (pesata da $\alpha_i \geq 0$), e il terzo termine è la "disutilità dell'essere avvantaggiato" (pesata da $\beta_i \geq 0$, tipicamente $\beta_i < \alpha_i$ — essere svantaggiati è peggio di essere avvantaggiati). Per $\alpha_i = \beta_i = 0$, si recupera la teoria classica dell'utilità monetaria.

Il modello di Fehr-Schmidt predice correttamente: la distribuzione delle offerte nell'UG (propositori offrono il 40–50% per anticipare il rifiuto dei rispondenti); il comportamento nel Dictator Game (donazioni positive anche in assenza di possibilità di rifiuto); il comportamento nel Public Goods Game (contributi positivi che declinano nel tempo e si mantengono più alti in presenza di punizione).

La struttura neurale del modello di inequity aversion localizza i parametri α_i e β_i in strutture neurali diverse: α_i (sensibilità alla disadvantageousness) è correlata con l'attivazione dell'insula anteriore e dell'amigdala; β_i (sensibilità alla advantageousness) è correlata con l'attivazione della corteccia cingolata anteriore e del caudato. Le lesioni selettive di queste aree producono comportamenti previsti dal modello: lesioni all'insula riducono il rifiuto delle offerte ingiuste (riduzione di α_i); lesioni al caudato riducono la punizione altruistica nei giochi pubblici (riduzione di β_i).

VI. I limiti del sistema decisionale: impulsività, addiction e patologie

VI.1 Il modello iperbolico dello sconto temporale e l'impulsività

Il discounting iperbolico — già introdotto nella sezione IV — ha implicazioni strutturali per il comportamento impulsivo. Un agente con discounting iperbolico ha funzione di valore:

$$V(A, t) = V_0(A) / (1 + k \cdot t)$$

dove $V_0(A)$ è il valore senza ritardo e k è il tasso di sconto. Per una scelta tra una ricompensa piccola-immediata (A: 10€ ora) e una ricompensa grande-ritardata (B: 15€ tra una settimana):

$$V(A, 0) = 10 \quad V(B, 0) = 15 / (1 + k \cdot 7)$$

Per $k = 0.5/\text{giorno}$: $V(B, 0) = 15 / 4.5 = 3.3\text{€} < 10\text{€} \rightarrow$ l'agente sceglie A (impulsivo). Per $k = 0.01/\text{giorno}$: $V(B, 0) = 15 / 1.07 = 14\text{€} > 10\text{€} \rightarrow$ l'agente sceglie B (paziente).

L'inversione della preferenza avviene quando entrambe le ricompense sono poste nel futuro: un agente con discounting iperbolico che oggi preferisce A (tra 1 settimana) su B (tra 2 settimane) potrebbe domani preferire B (tra 1 settimana) su A (immediata). Questa dinamica produce comportamenti come la procrastinazione (si pianifica di lavorare domani — ma quando "domani" arriva, si rimanda ancora).

Il substrato neurale del discounting iperbolico è localizzato nella competizione tra il sistema limbico-striato (NAcc, amigdala — sistema "myopic") e il sistema prefrontale dorsolaterale (DLPFC — sistema "patient"). McClure et al. (2004) hanno mostrato con fMRI che le ricompense immediate attivano preferenzialmente le strutture limbiche (NAcc, orbito-frontale mediale), mentre le ricompense ritardate attivano preferenzialmente la DLPFC. L'impulsività è associata a una relativa dominanza del sistema limbico sul prefrontale — la forma biologica del Sistema 1 che sopraffà il Sistema 2.

VI.2 Il cervello addicted: la sensitizzazione e il processo opponente

La dipendenza da sostanze è il caso estremo di patologia del sistema di valutazione. Solomon e Corbit (1974) proposero il modello del "processo opponente" per spiegare la struttura temporale dell'addiction: l'uso di una sostanza produce un processo primario a (euforia), che viene opposto da un processo b (disforia) di segno opposto. Con l'uso ripetuto, il processo b si rafforza (sensitizzazione del processo opponente) mentre il processo a si indebolisce (tolleranza). Il risultato è che la ricompensa soggettiva dalla sostanza diminuisce

progressivamente, mentre lo stato di baseline — in assenza della sostanza — diventa sempre più negativo.

La formalizzazione neurale di questo processo usa il modello dell'allostasi di Koob e Le Moal (2001). Sia $R(t)$ il livello di ricompensa soggettiva e $B(t)$ il baseline hedonic ("setpoint"):

$$R(t) = \text{effetto_droga}(t) - B(t)$$

Durante l'uso della droga:

$$dB/dt = \alpha \cdot \text{droga}(t)$$

Il baseline sale progressivamente con l'uso della droga — il "setpoint" edonico si sposta verso valori più alti, richiedendo dosi crescenti per produrre la stessa euforia (tolleranza) e producendo disforia (B elevato senza droga) in assenza della sostanza (astinenza).

Nel lungo termine:

$$B_{\text{addicted}} \gg B_{\text{normale}}$$

L'agente addicted non usa la droga per stare bene: la usa per stare "normalmente" — per ridurre il gap tra il baseline elevato e la baseline normale. La struttura si integra con la separazione wanting/liking di Berridge: il wanting rimane alto o aumenta (sensitizzazione dopaminergica) mentre il liking si riduce (tolleranza oppioidergica), producendo la ricerca compulsiva della droga senza piacere.

VII. La neuroscienza della moralità

VII.1 Il trolley problem e i circuiti neurali della moralità

Il trolley problem — nelle versioni di Foot (1967) e Thomson (1985) — è diventato il paradigma sperimentale per studiare i substrati neurali del ragionamento morale. Nella versione "footbridge" (il variante più critico), una persona deve decidere se spingere un uomo corpulento da un cavalcavia per fermare un tram in corsa che altrimenti ucciderà cinque persone. La maggioranza delle persone giudica questa azione moralmente sbagliata (anche se "razionalmente" produce il miglior esito consequenziale — 1 morte invece di 5).

Greene et al. (2001, 2004) hanno mostrato con fMRI che i dilemmi morali "personali" (che richiedono l'azione diretta danno a una persona specifica, come nella versione footbridge) attivano più fortemente le

aree associate alle emozioni sociali (giro frontale mediale, corteccia cingolata posteriore, sulco temporale superiore posteriore) rispetto ai dilemmi "impersonali" (che producono il danno in modo indiretto). I partecipanti che fanno scelte utilitariste nei dilemmi personali (spingono l'uomo) mostrano minore attivazione dell'MPFC e del cingolato posteriore — come se "sopprimer" la risposta emotiva per permettere il ragionamento consequenziale.

La struttura neurale corrisponde al modello duale: le emozioni sociali (il "non fare del male diretto") generano la risposta non-utilitarista immediata; il ragionamento consequenziale (il "massimizza le vite salvate") può sovrascriverla. La prevalenza della risposta deontologica (non spingere) nelle popolazioni normali riflette la forza relativa delle emozioni sociali rispetto al ragionamento consequenziale — un'asimmetria che ha radici evolutive: la risposta all'azione diretta su una persona è più "evolutive" (radicata nei meccanismi di empatia e aversione al danno) della risposta al calcolo consequenziale a lungo raggio.

VII.2 I limiti del cervello morale

Il cervello morale ha limiti strutturali che i filosofi spesso ignorano. Il bias di omissione (Monografia VI) produce giudizi morali asimmetrici tra azioni e omissioni — si attribuisce maggiore colpa a chi attiva attivamente un danno rispetto a chi omette di prevenirlo. Il myopic discounting rende difficile la valorizzazione appropriata delle conseguenze a lungo termine delle azioni morali (i danni futuri del cambiamento climatico sono sistematicamente sottopesati rispetto ai costi immediati della riduzione delle emissioni). L'in-group favoritism — la tendenza a dare peso morale maggiore ai membri del proprio gruppo — è radicata nelle strutture neurali dell'empatia ed è difficilmente correggibile attraverso il ragionamento esplicito.

Questi limiti del cervello morale non sono difetti correggibili: sono le manifestazioni biologiche delle stesse strutture che producono la cooperazione intragruppo, la protezione degli stessi, la risposta rapida alle minacce immediate. Il cervello è morale nel senso ecologico — adattato all'ambiente sociale del piccolo gruppo del Pleistocene — e produce distorsioni sistematiche nell'ambiente morale della modernità globale.

VIII. Conclusioni: la neurobiologia del valore come fondamento e limite

La neurobiologia della decisione ha prodotto la risposta biologica alla domanda filosofica di Aristotele: come scelgono gli agenti razionali? Il cervello implementa un sistema di apprendimento per rinforzo (TD learning nel sistema dopaminergico), un sistema di pianificazione model-based (via mesolimbica-prefrontale), un sistema di abitudini model-free (via nigrostriatale), e un sistema di valutazione sociale (OFC-insula-cingolato). Questi sistemi competono e si integrano, producendo scelte che sono razionali nel senso ecologico — ottimali nell'ambiente evolutivo — ma sistematicamente distorte nell'ambiente moderno.

I limiti strutturali del sistema decisionale biologico — il discounting iperbolico, il bias di omissione, l'in-group favoritism, la sensibilità alle emozioni immediate rispetto alle conseguenze future — non sono patologie correggibili ma proprietà necessarie di un sistema selezionato per un ambiente diverso. Comprendere questi limiti è comprendere la struttura dell'etica biologica — non per giustificarla, ma per progettare strutture sociali e istituzionali che li compensino, esattamente come la nudging compensa i limiti cognitivi sfruttandone invece di ignorandone la struttura.

Monografia XV

Il ragionamento controfattuale e la causalità nella cognizione biologica

Modelli causali, l'intervento come operazione formale, la simulazione mentale e i limiti dell'inferenza causale

I. La domanda che separa l'agente dal termostat

Un termostato risponde alla temperatura: quando la temperatura scende sotto la soglia, accende il riscaldamento; quando risale, lo spegne. Questo comportamento è finalistico — il termostato "persegue" il mantenimento della temperatura prefissata — e può essere descritto con lo stesso linguaggio teleologico che usiamo per gli agenti biologici. Ma nessuno direbbe che il termostato ragiona sulla temperatura: si limita a rispondere a essa.

La differenza tra il termostato e un agente che ragiona sull'ambiente non è nella complessità del comportamento — un termostato può essere arbitrariamente complesso — ma nella struttura dell'elaborazione interna: l'agente che ragiona mantiene un modello del mondo che include relazioni causali, usa quel modello per anticipare le conseguenze delle proprie azioni prima di eseguirle, e può deliberatamente considerare scenari controfattuali — stati del mondo che non si sono verificati ma che si sarebbero verificati se qualcosa fosse stato diverso. Il termostato non chiede mai "cosa sarebbe successo se avessi spento il riscaldamento prima?".

Il ragionamento controfattuale — la capacità di pensare a ciò che non è accaduto ma sarebbe potuto accadere — è considerato da molti ricercatori come la capacità cognitiva distintiva degli esseri umani rispetto agli altri primati. Suddendorf e Corballis (1997) la chiamano "mental time travel" (viaggio mentale nel tempo): la capacità di distaccarsi dal presente e simulare eventi passati e futuri. Però questa capacità non è tutta o niente: emerge gradualmente nello sviluppo, ha correlati neurali precisi, e — come questa monografia mostra — ha limiti strutturali che corrispondono ai limiti dell'inferenza causale come problema matematico.

II. La struttura formale della causalità: i modelli causali strutturali

II.1 La gerarchia causale di Pearl

Judea Pearl (2000, 2009) ha proposto una formalizzazione della causalità che distingue tre livelli di inferenza, ordinati in una gerarchia di potere esplicativo crescente:

Il primo livello è l'associazione (seeing): la capacità di osservare correlazioni tra variabili. Un agente al primo livello può rispondere a domande del tipo "qual è la probabilità di Y dato che $X = x$?" — può calcolare $P(Y \mid X = x)$ ma non può distinguere tra correlazione e causalità. Il termostato opera a questo livello.

Il secondo livello è l'intervento (doing): la capacità di predire le conseguenze di interventi — azioni deliberate che impongono un valore a una variabile indipendentemente dalle sue cause naturali. Un agente al secondo livello può rispondere a domande del tipo "qual è la probabilità di Y se faccio $X = x$?" — la notazione di Pearl è $P(Y \mid \text{do}(X = x))$, che differisce fondamentalmente da $P(Y \mid X = x)$. La differenza è la

struttura causale: $\text{do}(X = x)$ significa tagliare le frecce causali che entrano in X nel grafo causale e impostare $X = x$ forzatamente.

Il terzo livello è il controfattuale (imagining): la capacità di ragionare su ciò che sarebbe accaduto se qualcosa fosse stato diverso, condizionatamente a ciò che è effettivamente accaduto. Un agente al terzo livello può rispondere a domande del tipo "dato che $Y = y$ è accaduto (con $X = x$), cosa sarebbe accaduto a Y se avessi fatto $X = x'$ invece?" — la notazione è $P(Y_{\{X=x'\}} = y' \mid X = x, Y = y)$.

La gerarchia non è banale: ogni livello superiore richiede informazioni strutturali che non sono ricavabili dal livello inferiore. Non si può calcolare $P(Y \mid \text{do}(X))$ da dati osservazionali senza un modello causale (un grafo causale strutturale); non si può calcolare il controfattuale $P(Y_{\{X=x'\}} \mid X = x, Y = y)$ senza la struttura funzionale del modello causale (le equazioni strutturali, non soltanto il grafo).

II.2 I Structural Causal Models (SCM)

Un modello causale strutturale (SCM) di Pearl è una tripla $M = (U, V, F)$ dove:

$U = \{U_1, \dots, U_n\}$ sono le variabili esogene (i "rumori" — fattori non modellati che determinano la variabilità del sistema) $V = \{V_1, \dots, V_n\}$ sono le variabili endogene (le variabili del modello) $F = \{f_1, \dots, f_n\}$ sono le funzioni strutturali: $V_i = f_i(\text{pa}(V_i), U_i)$ dove $\text{pa}(V_i) \subset V$ sono i "genitori" di V_i nel grafo causale (le variabili che influenzano direttamente V_i).

L'operazione $\text{do}(V_i = v_i)$ viene implementata sostituendo la funzione f_i con la costante v_i — rimuovendo l'influenza dei genitori di V_i e imponendo il valore direttamente. Il grafo causale mutilato $M_{\{V_i = v_i\}} = (U, V, F')$ dove $F' = F$ con f_i sostituito da v_i .

I controfattuali richiedono un passo ulteriore: la "gemella" (twin world) di Pearl. Per calcolare $P(Y_{\{X=x'\}} \mid X = x, Y = y)$, si costruisce un mondo gemello M' identico a M ma con l'intervento $\text{do}(X = x')$. Il mondo gemello condivide le variabili esogene U con il mondo reale — la "stessa persona in un mondo alternativo" — ma le variabili endogene si propagano in modo diverso per effetto dell'intervento. Il valore controfattuale $Y_{\{X=x'\}}$ è il valore di Y nel mondo gemello.

Formalmente, per SCM lineari ($V = AV + BU$ con A matrice delle influenze causali e B matrice dei rumori):

$$V = (I - A)^{-1} B U \text{ (valori osservazionali)} \quad V^{\{\text{do}(X_i=x)\}} = (I - A_{\{-i\}})^{-1} B_{\{-i\}} U + E_i x_i \text{ (valori interventistici)}$$

dove A_{-i} è la matrice A con la i -esima riga azzerata (rimuoviamo le influenze sui genitori di X_i) e E_i è il vettore unitario nella direzione i (impostiamo $X_i = x_i$).

III. L'acquisizione del ragionamento causale nello sviluppo

III.1 La comprensione causale nel bambino

Gopnik e Schulz (2004) hanno mostrato che i bambini di 2–3 anni hanno già una capacità sofisticata di ragionamento causale che va al di là della semplice associazione: possono identificare le cause di eventi dall'osservazione di sequenze temporali, possono distinguere relazioni causali da relazioni statistiche spurie, e possono usare il ragionamento causale per pianificare interventi. Gopnik ha proposto che i bambini siano "natural scientists" — agenti che costruiscono e testano modelli causali del mondo con la struttura delle reti bayesiane.

La capacità di inferire strutture causali da dati osservazionali — il "causal learning" — è formalizzata attraverso il problema dell'identificabilità causale. Data una distribuzione di probabilità $P(V)$ sulle variabili V , è possibile identificare il grafo causale G che l'ha generata? In generale, no: la stessa distribuzione $P(V)$ è compatibile con più grafi causali (la "Markov equivalence class"). La discriminazione tra grafi causalmente equivalenti richiede dati interventistici (esperimenti) o assunzioni strutturali aggiuntive.

Gli esperimenti di Gopnik mostrano che i bambini comprendono intuitivamente la distinzione tra dati osservazionali e dati interventistici: "vedere che X accade insieme a Y " (correlazione) non implica "fare X produrrà Y " (causalità). Questa comprensione emerge intorno ai 3–4 anni attraverso lo sviluppo del reasoning by intervention — la capacità di ragionare su interventi deliberati come distincts dall'osservazione passiva.

III.2 La comprensione dei controfattuali nello sviluppo

La comprensione piena dei controfattuali — il terzo livello della gerarchia di Pearl — emerge più tardi nello sviluppo, intorno ai 4–5 anni, e coincide approssimativamente con l'acquisizione della teoria della mente (falsa credenza). Beck et al. (2006) hanno mostrato che la capacità di ragionare su controfattuali ("se A avesse fatto X invece di Y , cosa sarebbe successo?") è correlata con la capacità di rappresentare

falsa credenza — entrambe richiedono la simulazione di uno stato del mondo diverso da quello attuale.

Il parallelismo non è accidentale: la teoria della mente e il ragionamento controfattuale hanno la stessa struttura computazionale — entrambe richiedono la costruzione e la simulazione di un modello alternativo del mondo (un "mondo possibile") che differisce dal mondo attuale per il valore di alcune variabili. Nel caso della falsa credenza, la variabile è la credenza dell'agente target; nel caso del controfattuale, la variabile è la causa dell'evento di interesse.

La formalizzazione usa il framework degli SCM: la teoria della mente è l'inferenza dei parametri del SCM interno a un agente (il suo modello del mondo, le sue credenze e desideri), e il ragionamento controfattuale è il calcolo di controfattuali rispetto a quel SCM. Entrambe richiedono il terzo livello della gerarchia di Pearl.

IV. I correlati neurali del ragionamento causale e controfattuale

IV.1 La rete del default mode come simulatore

La rete del default mode (DMN) — il sistema di aree che si attivano a riposo e si disattivano durante compiti cognitivi focalizzati — ha proprietà che corrispondono esattamente a quelle di un simulatore di mondi alternativi. Buckner e Carroll (2007) hanno proposto che la DMN sia il substrato neurale del "prospective memory" e del "mental time travel" — non soltanto del ricordo del passato e dell'anticipazione del futuro, ma di qualsiasi forma di simulazione che richieda di distaccarsi dal presente.

Le aree principali della DMN — la corteccia prefrontale mediale (mPFC), il precuneo/corteccia cingolata posteriore (PCC/PCu), il giro angolare — mostrano una proprietà di co-attivazione: si attivano insieme in compiti che richiedono la simulazione di prospettive alternative (ragionamento controfattuale, prospettiva in prima persona di eventi futuri, comprensione di storie con prospettive di agenti diversi dal proprio). Questa co-attivazione suggerisce che la DMN implementa un sistema unificato di simulazione che è condiviso da tutte le forme di reasoning about alternatives.

Christoff et al. (2011) hanno mostrato che l'intensità dell'attivazione della DMN durante il ragionamento controfattuale è correlata positivamente con la creatività misurata con test divergenti — la

capacità di generare soluzioni alternative ai problemi. Questo correla la funzione della DMN come simulatore di mondi alternativi con la funzione creativa del cervello — la generazione di idee nuove è computazionalmente equivalente alla simulazione di stati del mondo controfattuali (stati che non esistono attualmente ma che potrebbero esistere).

IV.2 L'ippocampo come costruttore di scenari

Hassabis et al. (2007) hanno scoperto che pazienti con danni all'ippocampo mostrano deficit non soltanto nel ricordo episodico del passato ma anche nell'immaginazione di scenari futuri e nell'elaborazione di scenari controfattuali. Questo risultato — sorprendente perché si attendeva che l'ippocampo fosse coinvolto soltanto nella memoria, non nell'immaginazione — ha portato alla teoria della "costruzione di scene" (scene construction theory, Hassabis e Maguire, 2007): l'ippocampo è il substrato della capacità di costruire rappresentazioni spaziali coerenti di scenari complessi, indipendentemente dal fatto che quei scenari siano ricordi del passato, anticipazioni del futuro, o scenari controfattuali.

La formalizzazione usa il concetto di "schema epistemico" — una struttura generativa che permette al sistema di costruire rappresentazioni coerenti di scenari. L'ippocampo implementa questo schema attraverso la binding di elementi (persone, oggetti, luoghi) in pattern spaziali coerenti — la stessa funzione di binding analizzata nella Monografia I ma applicata agli scenari episodici invece che agli oggetti percettivi.

Il danno ippocampale produce scenari impoveriti e frammentati: i pazienti descrivono scene future o controfattuali con pochi dettagli spaziali, nessuna coesione narrativa, e incapacità di posizionare gli elementi dello scenario in relazione spaziale coerente. La "ricchezza" del ragionamento controfattuale dipende dalla capacità del sistema ippocampale di costruire scenari spazialmente strutturati — una dipendenza che lega il ragionamento astratto a una struttura neurale concretamente spaziale.

V. Il simulatore mentale e i suoi parametri

V.1 La simulazione come motore di inferenza

La visione della cognizione come simulazione è stata formalizzata da Hesslow (2002) nella "simulation hypothesis": il cervello usa le stesse

reti neurali per simulare azioni, percezioni ed emozioni che usa per eseguirle realmente. La simulazione è esecuzione inibitoria — il sistema esegue internamente i circuiti sensomotori ed emotivi, ma le risposte effettoriali vengono bloccate da meccanismi inibitori, permettendo al soggetto di "sperimentare" le conseguenze dell'azione senza eseguirla.

La formalizzazione nel framework del predictive coding è naturale: la simulazione è la propagazione top-down del modello generativo senza la correzione bottom-up dal segnale sensoriale. Nella percezione normale, il segnale sensoriale corregge le previsioni del modello; nella simulazione, il segnale sensoriale viene bloccato (o la sua precisione viene ridotta a zero), e le previsioni si propagano liberamente, producendo l'esperienza simulata.

Sia p_{true} la distribuzione sulla realtà sensoriale e p_{sim} la distribuzione sulla realtà simulata. La simulazione è:

$$p_{\text{sim}}(s) = p_{\text{model}}(s \mid \text{scenario_simulato})$$

senza la correzione bayesiana del segnale sensoriale esterno. La differenza tra percezione e simulazione è la presenza/assenza del segnale di correzione:

$$\text{Percezione: } p_{\text{post}}(s) \propto p_{\text{sensor}}(y \mid s) \cdot p_{\text{model}}(s) \quad \text{Simulazione: } p_{\text{sim}}(s) = p_{\text{model}}(s \mid \text{scenario})$$

La simulazione produce stati interni che sono simili agli stati percettivi (stesse reti neurali attivate) ma con precisione ridotta e controllabilità superiore — il soggetto può "dirigere" la simulazione in modi che non può dirigere la percezione.

V.2 I parametri della simulazione: fedeltà, risoluzione e orizzonte temporale

La simulazione mentale come sistema computazionale ha tre parametri fondamentali che ne determinano la qualità e i limiti.

La fedeltà della simulazione è la corrispondenza tra gli stati simulati e gli stati reali che si produrrebbero se lo scenario simulato si verificasse. Una simulazione ad alta fedeltà produce previsioni accurate delle conseguenze delle azioni; una a bassa fedeltà produce previsioni distorte. La fedeltà dipende dalla qualità del modello generativo — dalla precisione con cui il modello generativo del cervello cattura le dinamiche causali del mondo reale.

I limiti della fedeltà emergono dalla struttura del modello generativo: il modello è necessariamente incompleto (non modella tutte le variabili

rilevanti) e approssimato (le relazioni causali sono semplificate rispetto alla realtà). Le simulazioni di eventi lontani dall'esperienza personale — situazioni mai incontrate nella propria vita — hanno fedeltà bassa: il cervello deve estrapolare il modello generativo oltre il dominio per cui è stato calibrato.

La risoluzione temporale della simulazione è la granularità temporale degli stati simulati. Simulazioni a breve termine (pochi secondi) hanno alta risoluzione — possono includere dettagli spaziali e motori; simulazioni a lungo termine (mesi, anni) hanno bassa risoluzione — si riducono a rappresentazioni schematiche degli stati di interesse senza i dettagli intermedi. La risoluzione decresce approssimativamente come:

$$R(T) \propto 1/T$$

dove T è l'orizzonte temporale della simulazione. Questo produce un'asimmetria sistematica nella valutazione degli esiti temporalmente vicini vs. lontani — una fonte biologica del discounting temporale analizzato nella Monografia XIV.

L'orizzonte temporale è il limite massimo di futuro che può essere simulato in modo coerente. Per gli esseri umani, l'orizzonte temporale del ragionamento esplicito è dell'ordine di alcuni anni per le decisioni pratiche, alcune decadi per le pianificazioni a lungo termine, e si dissolve progressivamente per orizzonti più lunghi. I processi che si svolgono su scale di secoli o millenni (cambiamento climatico, evoluzione tecnologica, degenerazione demografica) sfuggono all'orizzonte del simulatore mentale e non possono essere "vissuti" come urgenti attraverso la simulazione — una delle ragioni per cui le crisi a lungo termine come il cambiamento climatico non evocano la stessa risposta emotiva delle crisi immediate.

VI. I limiti formali dell'inferenza causale

VI.1 Il problema dell'identificabilità causale

Il problema fondamentale dell'inferenza causale è il seguente: dato un insieme di variabili V e una distribuzione osservazionale $P(V)$, è possibile identificare la struttura causale G (il grafo causale che ha generato $P(V)$) e i parametri quantitativi del SCM? In generale, la risposta è no — il problema è sottodeterminato.

Il concetto di equivalenza di Markov formale: due grafi causali G e G' sono Markov equivalenti se producono le stesse d-separazioni — le

stesse strutture di indipendenza condizionale nella distribuzione $P(V)$. Grafi Markov equivalenti non sono distinguibili da dati osservazionali soli, indipendentemente dalla quantità di dati disponibili.

Per distinguere tra grafi Markov equivalenti è necessario o (a) eseguire esperimenti — interventi $do(V_i = v_i)$ che rompono alcune equivalenze — o (b) imporre assunzioni strutturali sul modello (assunzioni di causalità temporale, assunzioni di funzioni lineari, assunzioni di errori non gaussiani).

Questo limite è fondamentale per il ragionamento causale biologico: il cervello non ha accesso illimitato a esperimenti — non può sistematicamente fare $do(V_i = v_i)$ per tutte le variabili rilevanti nella propria vita. Deve inferire strutture causali principalmente da dati osservazionali, con le ambiguità strutturali che questo comporta. Le distorsioni cognitive nella percezione della causalità — il bias di hind sight ("l'avevo detto"), il bias di attribuzione causale (attribuire gli eventi positivi a cause interne e quelli negativi a cause esterne), la confusione tra correlazione e causalità — sono manifestazioni biologiche dei limiti dell'inferenza causale dai dati osservazionali.

VI.2 La complessità computazionale dell'inferenza causale

L'inferenza causale ottimale — il calcolo del grafo causale che massimizza la verosimiglianza dei dati osservazionali tra tutti i grafi causali possibili — è computazionalmente intrattabile in generale. Il problema di trovare il DAG (directed acyclic graph) che massimizza un criterio di scoring sulle variabili V è NP-hard (Chickering, 1996): il numero di DAG possibili su n variabili cresce super-esponenzialmente (più che $2^{(n^2)}$), e nessun algoritmo può esplorare l'intero spazio dei DAG in tempo polinomiale.

Questo significa che il cervello non può implementare l'inferenza causale ottimale in senso stretto: deve usare euristiche che approssimano l'inferenza ottimale in modo computazionalmente trattabile. Le euristiche identificate nella letteratura psicologica sulla percezione della causalità — la prossimità temporale (le cause precedono gli effetti di poco), la covariazione (le cause co-variano con gli effetti), la contiguità spaziale (le cause sono spazialmente vicine agli effetti) — sono regole locali che approssimano l'inferenza causale bayesiana nel regime in cui le relazioni causali sono semplici e i ritardi temporali sono brevi.

Queste euristiche sono sistematicamente distorte quando le relazioni causali violano le assunzioni che le sottendono: cause remote (lontane nel tempo o nello spazio), cause multiple con effetti interattivi, effetti

ritardati. Le patologie del ragionamento causale umano — il pensiero magico, la superstizione, la post hoc ergo propter hoc — sono le conseguenze prevedibili dell'applicazione di euristiche calibrate su relazioni causali semplici a relazioni causali complesse.

VII. Il rimpianto come computazione controfattuale

VII.1 La struttura del rimpianto

Il rimpianto — l'emozione che emerge dal confronto tra il risultato ottenuto e il risultato che si sarebbe ottenuto con una scelta diversa — è computazionalmente una valutazione controfattuale di terzo livello nella gerarchia di Pearl. "Se avessi preso la strada diversa, sarei arrivato in tempo" è un controfattuale: confronta lo stato reale (ritardo) con lo stato che si sarebbe verificato in un mondo gemello (arrivo in tempo) sotto l'intervento controfattuale (prendere la strada diversa).

Zeelenberg e Pieters (2007) hanno mostrato che il rimpianto influenza sistematicamente le future decisioni: individui che hanno sperimentato rimpianto per un'omissione passata (non aver comprato un'azione) tendono a comportarsi in modo più attivo nelle decisioni future; individui che hanno sperimentato rimpianto per un'azione passata (aver comprato un'azione sbagliata) tendono a comportarsi in modo più passivo. Il rimpianto calibra il sistema decisionale aggiornando i parametri del modello causale interno — aggiunge informazione controfattuale al modello.

La neurobiologia del rimpianto localizza il segnale controfattuale nella corteccia orbitofrontale mediale (mOFC) e nel cingolato anteriore (ACC). Coricelli et al. (2005) hanno mostrato con fMRI che la risposta neurale al rimpianto (confronto tra esito reale e migliore alternativa non scelta) produce attivazione della mOFC che scala con la magnitudine del rimpianto — proporzionale alla differenza tra esito reale e esito controfattuale. L'attivazione è più forte per rimpianti da azione che da omissione, in accordo con il bias di commissione della Monografia VI.

La formalizzazione usa la teoria del "regret" di Loomes e Sugden (1982): la funzione di utilità modificata dal rimpianto è:

$$V_{\text{regret}}(a, \omega) = U(a, \omega) + R(U(a, \omega) - U(a^*, \omega))$$

dove $U(a, \omega)$ è l'utilità dell'azione a nell'esito ω , $U(a^*, \omega)$ è l'utilità dell'azione ottimale nello stesso esito ω (il controfattuale), e R è la

funzione di rimpianto (negativa quando $U(a) < U(a^*)$, zero altrimenti). Il rimpianto trasforma la valutazione delle azioni da assoluta (basata sull'utilità dell'esito) a relativa (basata sul confronto con l'alternativa migliore non scelta).

VIII. La causalità e il ragionamento morale

VIII.1 Il principio del doppio effetto e la struttura causale

Il principio del doppio effetto — un principio etico medievale formalizzato da Tommaso d'Aquino e ancora influente nella bioetica moderna — afferma che è moralmente lecito eseguire un'azione che ha sia conseguenze buone che cattive se: (1) l'azione in sé non è intrinsecamente malvagia, (2) l'intento è produrre la conseguenza buona, (3) la conseguenza cattiva non è il mezzo per produrre quella buona, e (4) c'è proporzione tra il bene e il male prodotti.

La condizione (3) — la conseguenza cattiva non deve essere il mezzo per produrre quella buona — è una condizione causale: nel grafo causale dello scenario, deve esserci un percorso causale dall'azione alla conseguenza buona che non passa attraverso la conseguenza cattiva. Nel trolley problem "footbridge" (spingere la persona corpulenta), la conseguenza buona (le cinque vite salvate) si produce causalmente attraverso la conseguenza cattiva (la morte del corpulento): la persona serve come mezzo. Nel trolley problem "switch" (deviare il tram), la conseguenza buona si produce attraverso la deviazione del tram, e la morte del singolo è una conseguenza laterale, non il mezzo.

La struttura causale — non la struttura consequenziale (lo stesso numero di morti in entrambi i casi) — è ciò che distingue moralmente i due scenari nel giudizio intuitivo. Questo conferma che il ragionamento morale umano è sensibile alla struttura causale degli scenari, non soltanto ai risultati: il sistema morale biologico implementa qualcosa di simile al principio del doppio effetto come euristico causale.

Mikhail (2007) ha formalizzato questa intuizione nella "moral grammar" — una grammatica universale della moralità analoga alla grammatica universale di Chomsky per il linguaggio: gli esseri umani applicano strutture causali e intenzionali nell'analisi degli scenari morali, producendo giudizi che corrispondono alle condizioni del principio del doppio effetto anche senza averlo mai formulato esplicitamente. Questa

grammatica morale è automatica, inconscia, e condivisa cross-culturalmente — una tesi che ha trovato supporto empirico parziale in studi su popolazioni diverse (Hauser et al., 2007).

IX. L'emergenza della causalità da strutture di co-variazione

IX.1 Il cervello come rivelatore causale bayesiano

La domanda fondamentale della neuroscienza causale è: come il cervello inferisce strutture causali dall'esperienza? Il framework bayesiano per l'apprendimento causale (Griffiths e Tenenbaum, 2005) propone che il cervello mantenga distribuzioni di probabilità su strutture causali alternative e le aggiorni con l'evidenza in modo bayesiano ottimale.

La distribuzione prior sulle strutture causali $P(G)$ riflette le assunzioni generali del sistema sulla struttura del mondo (ad esempio la preferenza per strutture causali semplici, la preferenza per la causalità temporalmente ordinata, il prior contro le cause multiple indipendenti). La verosimiglianza $P(D | G)$ è la probabilità di osservare i dati D dato che la struttura causale è G . L'aggiornamento bayesiano produce:

$$P(G | D) \propto P(D | G) \cdot P(G)$$

La struttura causale più probabile date le osservazioni è la moda di questa distribuzione posteriore.

Griffiths e Tenenbaum (2005) mostrano che per compiti di apprendimento causale semplici (un singolo candidato causa, un singolo effetto), gli esseri umani si comportano in accordo con questo framework bayesiano. Le deviazioni sistematiche osservate in compiti più complessi (cause multiple, effetti ritardati, feedback loops) corrispondono alle approssimazioni euristiche che il sistema cognitivo usa quando il calcolo bayesiano esatto è computazionalmente troppo costoso.

X. Conclusioni: la simulazione controfattuale come confine del vivente

Questa monografia chiude la pars costruens del ciclo con una riflessione sulla natura del ragionamento controfattuale come confine

tra i sistemi che reagiscono al mondo e i sistemi che modellano il mondo. Il termostato reagisce; l'agente dotato di modelli causali anticipa, simula, controfattualizza. La capacità di pensare a ciò che non è accaduto — di costruire mondi gemelli dove le cose sarebbero andate diversamente — è la base della pianificazione, dell'apprendimento dagli errori, della creatività, della moralità e della scienza.

Ma questa capacità ha limiti strutturali che la matematica dell'inferenza causale rende precisi: i limiti dell'identificabilità causale (lo stesso set di dati è compatibile con strutture causali diverse), della complessità computazionale (l'inferenza causale ottimale è NP-hard), della fedeltà della simulazione (i modelli generativi sono incompleti), e dell'orizzonte temporale (la risoluzione della simulazione decade con la distanza temporale).

Questi limiti non sono difetti correggibili — sono proprietà strutturali di qualsiasi sistema che inferisce la causalità da dati finiti in tempo finito. Comprendere questi limiti è il prerequisito per comprendere la struttura del ragionamento controfattuale biologico, le sue distorsioni sistematiche, e le condizioni in cui può essere migliorato attraverso strumenti formali — le metodologie della statistica causale che permettono di superare, almeno parzialmente, i limiti cognitivi dell'inferenza causale intuitiva.

Con questa monografia si chiude la pars costruens del ciclo. Il terzo movimento affronta ora le sintesi: come lo sviluppo costruisce il cervello, come la mente bayesiana emerge dall'evoluzione, come la semantica nasce dal segnale biologico, e come il corpo — nella sua irriducibile fisicità — pone il confine ultimo di ogni teoria della mente.

Monografia XVI

L'algorithmo dello sviluppo

Evo-devo, reti di regolazione genica, la grammatica dello spazio corporeo e i limiti della riprogrammazione

I. Il paradosso dell'uguaglianza genetica

Il genoma umano contiene circa 20.000–25.000 geni codificanti per proteine. Un neurone della corteccia prefrontale e una cellula del fegato hanno esattamente lo stesso genoma — eppure differiscono in modo

straordinario nella morfologia, nella funzione, nel metabolismo e nel ciclo vitale. Come la stessa informazione genetica produce strutture così diverse? E — problema ancora più profondo — come quella stessa informazione produce, in due mesi di sviluppo embrionale, la transizione da una singola cellula fertilizzata a un organismo con miliardi di cellule organizzate in centinaia di tipi cellulari distinti, con topologia corporea precisa e funzione coordinata?

Questa domanda — la domanda fondamentale della biologia dello sviluppo — ha ricevuto risposta parziale nei decenni successivi alla scoperta della struttura del DNA, ma la sua risposta completa coinvolge meccanismi che trascendono la semplice codifica proteica: le reti di regolazione genica, i codici epigenetici, i gradienti morfogenetici, la meccanobiologia cellulare. Questo quarto livello di organizzazione biologica — tra il genoma (l'informazione) e l'organismo (la struttura) — è il tema di questa monografia.

Il campo della biologia evolutiva dello sviluppo (evo-devo) ha rivelato negli ultimi trent'anni che il piano corporeo degli animali — la struttura dei segmenti del corpo, la posizione degli organi, la topologia dei sistemi — è controllato da un piccolo insieme di geni altamente conservati attraverso tutta la vita animale. Gli stessi geni Hox che determinano la struttura delle vertebre dell'embrione umano determinano la struttura dei segmenti dell'embrione della *Drosophila*. Questa conservazione profonda rivela che l'algoritmo dello sviluppo non è stato inventato separatamente per ogni lineaggio animale: è stato inventato una volta, in un antenato comune di tutti i bilateri, e conservato per 550 milioni di anni attraverso tutte le rivoluzioni morfologiche.

II. Le reti di regolazione genica: struttura e dinamica

II.1 La struttura formale delle GRN

Una rete di regolazione genica (Gene Regulatory Network, GRN) è un grafo orientato $G = (V, E)$ dove V è l'insieme dei geni (o dei prodotti genetici — mRNA, proteine, RNA non codificanti) e E è l'insieme delle interazioni regolatorie: archi diretti da un gene regolatore al gene bersaglio, con segno positivo (attivazione) o negativo (repressione). Ogni nodo i della GRN ha un'equazione di stato:

$$dx_i/dt = f_i(x_{\{pa(i)\}}) - \delta_i \cdot x_i$$

dove x_i è il livello di espressione del gene i , $pa(i)$ sono i regolatori upstream di i (i "genitori" nel grafo), f_i è la funzione di regolazione (tipicamente non lineare — sigmoide di Hill per la cooperatività), e δ_i è il tasso di degradazione.

La funzione di Hill per un'attivazione con cooperatività n è:

$$f_i(x_j) = V_{\max} \cdot x_j^n / (K^n + x_j^n)$$

dove K è la concentrazione di metà attivazione e n è il coefficiente di Hill (cooperatività). Per $n = 1$, la risposta è iperbolica (non cooperativa); per $n \geq 2$, la risposta è sigmoideale (cooperativa) con una transizione sempre più netta al crescere di n . La cooperatività produce bistabilità: il sistema può avere due stati stazionari stabili — uno con alta espressione e uno con bassa espressione — separati da uno stato instabile.

La bistabilità è il meccanismo di base della determinazione cellulare: una cellula staminale pluripotente è in uno stato bistabile, e segnali di differenziazione la "commutano" verso uno dei possibili stati terminali (epatica, neurale, muscolare) attraverso la stabilizzazione di uno dei minimi del paesaggio epigenetico di Waddington.

II.2 Il paesaggio epigenetico di Waddington

Conrad Waddington (1957) propose la metafora del "paesaggio epigenetico" per descrivere il processo di differenziazione cellulare: immaginare il destino di una cellula come una pallina che rotola su un paesaggio collinare, con le valli (gli attractor states) che corrispondono ai tipi cellulari terminali e le colline (i punti di sella e le creste) che corrispondono agli stati transitori e alle biforcazioni tra diversi destini.

La formalizzazione matematica del paesaggio di Waddington usa il potenziale quasipotenziale $U(x)$ di un sistema dinamico stocastico. Per un sistema di equazioni differenziali stocastiche:

$$dx = f(x) dt + \sigma dW$$

dove $f(x)$ è il campo vettoriale deterministico, σ^2 è l'intensità del rumore, e dW è il rumore bianco, il quasipotenziale è definito come:

$$U(x^*) = -\lim_{\sigma \rightarrow 0} \{ \sigma^{-2} \log P_{ss}(x^*) \}$$

dove $P_{ss}(x^*)$ è la densità stazionaria della distribuzione nello stato x^* . I minimi di U corrispondono agli stati stazionari stabili (i tipi cellulari); le selle di U corrispondono ai punti di biforcazione tra destini alternativi.

La velocità di transizione tra stati (differenziazione cellulare spontanea) è governata dalla legge di Kramers:

$$k_{\{A \rightarrow B\}} \propto \exp(-(U_{\text{sella}} - U_A) / \sigma^2)$$

dove U_A è il potenziale nel bacino A e U_{sella} è il potenziale alla sella. Perturbazioni esterne (segnali di differenziazione) abbassano le barriere tra certi bacini, facilitando la transizione.

II.3 I "nodi master" e la gerarchia regolatoria

Le GRN biologiche non sono reti piatte — hanno struttura gerarchica con "nodi master" (master regulators) che controllano interi programmi di espressione genetica. MyoD è il master regulator della differenziazione muscolare: la sua espressione ectopica (l'espressione in cellule non muscolari) è sufficiente a ri-programmare cellule di qualsiasi tipo verso il destino muscolare. Pax6 è il master regulator dell'occhio: la sua espressione ectopica nell'ala della *Drosophila* induce la formazione di strutture oculari funzionali.

La struttura formale dei nodi master usa la teoria delle reti multi-strato. Sia H_1 il primo livello della GRN (i signaling genes che ricevono i segnali ambientali), H_2 il secondo livello (i geni regolatori che integrano i segnali), e H_3 il terzo livello (i geni effettori che producono le proteine strutturali e funzionali). Le interazioni sono:

$$H_1 \rightarrow H_2 \rightarrow H_3$$

con i nodi master in H_2 che coordinano l'espressione di interi moduli di H_3 . Un master regulator $m \in H_2$ controlla un modulo $M \subset H_3$ di geni co-regolati:

$$\forall g \in M: x_g(t) \approx \varphi_m(x_m(t)) \cdot \text{baseline}_g$$

dove φ_m è la funzione di attivazione del master regulator m e baseline_g è il livello basale del gene g . L'attivazione di m "accende" l'intero modulo M coordinatamente — l'intera "cassetta" del programma muscolare o dell'occhio o della notocorda.

III. I geni Hox e la grammatica del corpo

III.1 La struttura del complesso Hox

I geni Hox — omeotici, ossia che determinano l'identità delle strutture corporee lungo l'asse antero-posteriore — costituiscono il sistema di coordinate spaziali dell'asse corporeo negli animali bilaterali. Negli insetti (*Drosophila*) vi sono 8 geni Hox organizzati in un singolo

cluster; nei vertebrati vi sono 39 geni Hox organizzati in quattro cluster (HoxA, HoxB, HoxC, HoxD). La colinearità è la proprietà fondamentale di questi cluster: l'ordine fisico dei geni sul cromosoma corrisponde all'ordine delle regioni corporee in cui sono espressi — i geni a 3' del cluster sono espressi nelle regioni anteriori, quelli a 5' nelle regioni posteriori.

La colinearità è una delle regolarità più sorprendenti della biologia dello sviluppo: l'ordine lineare dell'informazione nel cromosoma (la sequenza genomica) si mappa sull'ordine spaziale delle strutture corporee (la posizione antero-posteriore). Questo non è accidentale: l'ordine genomico riflette l'ordine temporale dell'attivazione durante lo sviluppo (i geni 3' vengono espressi prima dei geni 5', nella stessa sequenza antero-posteriore), e questa temporalità è conservata perché la struttura cromatinica che regola l'apertura sequenziale dei cluster Hox dipende dall'ordine fisico dei geni.

Il sistema Hox è formalmente analogo a un codice d'indirizzo spaziale: ogni posizione lungo l'asse antero-posteriore del corpo embrionale è caratterizzata da una "combinazione Hox" — il profilo di espressione dei geni Hox in quella posizione. La combinazione Hox determina l'identità della struttura che si svilupperà in quella posizione: una combinazione di bassa espressione di geni posteriori corrisponde a un segmento cervicale della colonna vertebrale; un'alta espressione dei geni posteriori corrisponde a un segmento lombare. Le mutazioni nei geni Hox producono omeosi — la sostituzione di una struttura con un'altra (classico esempio: la mutazione *Antennapedia* in *Drosophila* che produce la crescita di un'anca dove dovrebbe crescere un'antenna).

III.2 La grammatica corporea come sistema generativo

Il parallelismo tra la grammatica Hox e la grammatica linguistica non è puramente analogico — ha una struttura formale precisa. Come una grammatica formale genera un insieme di stringhe attraverso l'applicazione ricorsiva di regole di riscrittura, il sistema di regolazione dello sviluppo genera la struttura corporea attraverso l'applicazione ricorsiva di programmi di differenziazione.

In notazione formale, sia Σ l'alfabeto dei tipi cellulari possibili e sia Γ l'insieme dei programmi di sviluppo (le GRN che trasformano un tipo cellulare in altri tipi). Lo sviluppo è una grammatica di L-System (Lindenmayer, 1968) — una grammatica parallela che riscrive tutti i simboli contemporaneamente in ogni passo:

$S \rightarrow \text{derivazione} \rightarrow \text{struttura corporea}$

dove S è la cellula uovo fertilizzata (il simbolo iniziale), ogni passo di derivazione corrisponde a una divisione cellulare accompagnata dalla differenziazione (la riscrittura), e la struttura corporea finale è la "stringa" terminale della grammatica.

La struttura dell'L-System è:

$$G = (V, P, \omega)$$

dove V è l'insieme dei tipi cellulari (variabili), P è l'insieme delle regole di produzione (programmi di sviluppo: $\text{tipo_cellulare} \rightarrow \{\text{figli}\}$), e ω è l'assioma (la cellula uovo). La derivazione produce la struttura corporea come un albero di divisioni cellulari con annotazioni di tipo cellulare.

La colinearità Hox è il meccanismo che impone l'ordine spaziale alla grammatica: assicura che la sequenza di attivazioni dei programmi di sviluppo segua la topologia antero-posteriore del corpo, non producendo strutture topologicamente incoerenti.

IV. La regolazione epigenetica come livello intermedio

IV.1 La struttura del codice epigenetico

L'epigenetica — nello stretto senso molecolare contemporaneo — è lo studio delle modifiche covalenti del DNA e degli istoni che influenzano l'espressione genica senza alterare la sequenza del DNA. Le modifiche epigenetiche principali sono:

La metilazione del DNA: aggiunta di un gruppo metile ($-\text{CH}_3$) alla citosina nelle isole CpG. La metilazione del promotore silenzia tipicamente il gene — la trascrizione è impedita dall'impossibilità della RNA polimerasi di legarsi al promotore metilato.

Le modifiche istoniche: acetilazione, metilazione, fosforilazione, ubiquitinazione delle code istoniche. L'acetilazione dei lisine istonici (H3K9ac, H3K27ac) è associata all'attivazione della trascrizione; la metilazione di H3K27me3 è associata al silenziamento; la metilazione di H3K4me3 è associata all'attivazione. La combinazione delle modifiche istoniche su un gene forma il "codice istonico" (Strahl e Allis, 2000) — un codice combinatoriale che regola la trascrizione con finezza superiore alla semplice presenza/assenza del gene.

La struttura della cromatina: l'apertura o chiusura della cromatina (eucromatina vs. eterocromatina) determina l'accessibilità dei geni alla macchinaria trascrizionale. La cromatina aperta permette il legame dei fattori di trascrizione; la cromatina chiusa blocca la trascrizione anche in presenza di fattori di trascrizione attivi.

Il codice epigenetico totale per una cellula è una matrice $C \in \{\text{metilato, non-metilato}\}^N \times \{\text{stato_istonico}\}^M \times \{\text{aperta, chiusa}\}^L$ — un oggetto ad alta dimensionalità che specifica la "configurazione regolatoria" dell'intero genoma nella cellula. Il tipo cellulare di una cellula è determinato, oltre che dal genotipo, dalla configurazione epigenetica — diversi tipi cellulari con lo stesso genotipo hanno diverse configurazioni epigenetiche.

IV.2 La trasmissione epigenetica e i suoi limiti

La caratteristica più sorprendente del codice epigenetico è la sua trasmissibilità: quando una cellula si divide, la configurazione epigenetica viene (parzialmente) trasmessa alle cellule figlie. Questo significa che la "memoria" del tipo cellulare — l'identità di una cellula epatica o neurale — è mantenuta non soltanto dal genotipo ma anche dall'epigenoma, ed è trasmessa alle cellule figlie attraverso la mitosi.

Il meccanismo di trasmissione della metilazione del DNA è preciso: la DNA metiltransferasi di mantenimento DNMT1 ri-metila il filamento del DNA neo-sintetizzato di fronte al filamento parentale già metilato, "copiando" la metilazione al sito originale. Questo meccanismo è semi-conservativo, come la replicazione del DNA stesso, con una fedeltà di trasmissione di circa 95–99% per sito per divisione.

Il limite della trasmissione epigenetica è la velocità di deriva epigenetica — l'accumulo di errori nella trasmissione del codice epigenetico nel corso delle divisioni cellulari. Sia p_{err} la probabilità di errore per sito per divisione ($p_{\text{err}} \approx 0.01\text{--}0.05$). Dopo n divisioni, la proporzione di siti correttamente trasmessi è:

$$P_{\text{corretto}}(n) = (1 - p_{\text{err}})^n$$

Per $p_{\text{err}} = 0.02$ e $n = 50$ divisioni: $P_{\text{corretto}} \approx (0.98)^{50} \approx 0.36$ — soltanto il 36% dei siti epigenetici è ancora correttamente metilato dopo 50 divisioni. Questo è il meccanismo del "epigenetic aging" — la deriva epigenetica progressiva nel corso della vita che contribuisce all'invecchiamento cellulare e all'aumento del rischio di cancro.

V. La plasticità fenotipica e la norma di reazione

V.1 La norma di reazione come funzione genotipo-fenotipo

La norma di reazione (Woltereck, 1909; Schmalhausen, 1949) è la funzione che mappa il genotipo G nei fenotipi prodotti in ambienti diversi E :

$$NR: G \times E \rightarrow P$$

Per un dato genotipo G , la norma di reazione $NR(G, \cdot) : E \rightarrow P$ descrive come il fenotipo varia al variare dell'ambiente. La plasticità fenotipica è la variabilità del fenotipo in risposta all'ambiente — sistemi con alta plasticità producono fenotipi molto diversi in ambienti diversi; sistemi con bassa plasticità producono fenotipi simili indipendentemente dall'ambiente.

La selezione naturale opera su fenotipi, non su genotipi. Poiché lo stesso genotipo produce fenotipi diversi in ambienti diversi, la relazione tra selezione e genotipo è mediata dalla norma di reazione — e la stessa selezione può avere effetti molto diversi su genotipi con norme di reazione diverse. Questo è il problema della " $G \times E$ interaction" (interazione genotipo-ambiente): l'effetto fenotipico di un genotipo dipende dall'ambiente, e viceversa.

La plasticità fenotipica adattiva — la capacità di produrre fenotipi adeguati in ambienti diversi — è essa stessa un carattere selezionato. Sistemi con alta plasticità adattiva possono occupare range ambientali più ampi; sistemi specializzati (bassa plasticità) possono essere più efficienti nei propri ambienti ottimali ma hanno range ambientali più stretti.

V.2 L'assimilazione genetica: dalla plasticità alla canalization

Conrad Waddington (1953) ha mostrato sperimentalmente che caratteri fenotipici prodotti inizialmente come risposte plastiche all'ambiente possono, nel corso dell'evoluzione, diventare costitutivi — espressi indipendentemente dall'ambiente. Questo processo — l'assimilazione genetica — è il meccanismo evolutivo attraverso cui la plasticità si "canalize" in un fenotipo fisso.

Il meccanismo formale dell'assimilazione genetica usa il concetto di canalizzazione di Waddington: il processo sviluppativo di un carattere può essere più o meno "canalizzato" — più o meno robusto alle perturbazioni. Un carattere altamente canalizzato produce lo stesso fenotipo nonostante variazioni genetiche e ambientali; un carattere poco canalizzato produce fenotipi diversi in risposta a piccole perturbazioni.

La selezione che agisce sul fenotipo canalizzato ha effetti indiretti sul genotipo: seleziona per i geni che aumentano la canalizzazione di quel fenotipo, rendendo la sua espressione sempre meno dipendente dall'ambiente specifico in cui era inizialmente indotta. Nel limite, il fenotipo diventa costitutivo — espresso in tutti gli ambienti indipendentemente dallo stimolo originale.

Formalmente, sia x il valore del fenotipo e sia T la soglia per l'espressione del fenotipo nell'ambiente ambiguo (il fenotipo è espresso se $x > T$). L'assimilazione genetica è la riduzione di T nel corso dell'evoluzione verso $T \rightarrow 0$ (il fenotipo è espresso indipendentemente dall'ambiente) — una selezione direzionale sulla distribuzione di x nella popolazione.

VI. I moduli di sviluppo e l'evolubilità

VI.1 La struttura modulare dello sviluppo

L'evolubilità — la capacità di un sistema biologico di generare variazione fenotipica utile attraverso la variazione genetica — dipende criticamente dalla struttura modulare dello sviluppo. Un sistema di sviluppo altamente integrato (in cui tutti i geni interagiscono con tutti gli altri) ha bassa evolubilità: qualsiasi variazione genetica produrrebbe effetti pleiotropici globali, e quasi tutte le variazioni sarebbero deleterie. Un sistema modulare (in cui i geni sono organizzati in moduli semi-indipendenti) ha alta evolubilità: variazioni all'interno di un modulo non perturbano gli altri moduli, permettendo un'esplorazione più fine del paesaggio fenotipico.

Clune, Mouret e Lipson (2013) hanno mostrato formalmente attraverso simulazioni evolutive che la selezione in ambienti variabili (che richiedono la ri-ottimizzazione di parti diverse del fenotipo indipendentemente) produce spontaneamente struttura modulare nelle reti regolatorie. Il meccanismo è diretto: sistemi con struttura modulare possono ri-ottimizzare un modulo in risposta a cambiamenti ambientali in quel modulo senza disturbare gli altri moduli — un vantaggio selettivo che porta all'emergenza e al mantenimento della modularità.

La struttura modulare delle GRN ha una misura formale analoga alla modularità dei grafi della Monografia X:

$$Q_{GRN} = \sum_m [L_m/L - (d_m/2L)^2]$$

dove L_m è il numero di interazioni all'interno del modulo m , L è il numero totale di interazioni, e d_m è la somma dei gradi dei nodi nel modulo m . Una GRN con alta modularità Q_{GRN} ha poca pleiotropia inter-modulo (le perturbazioni di un modulo non si propagano agli altri) e quindi alta evolvibilità.

VII. I limiti dell'algoritmo dello sviluppo: le vincolature developmental

VII.1 Le vincolature sviluppative come impossibilità morfologiche

Non tutti i fenotipi concepibili sono realizzabili nello sviluppo: alcune morfologie non possono essere prodotte da nessun percorso sviluppativo realistico a partire da nessun genotipo biologico. Queste impossibilità — le "developmental constraints" di Gould e Lewontin (1979, discussi nella Monografia II) — limitano lo spazio dei fenotipi realizzabili e quindi lo spazio di variazione disponibile all'evoluzione.

Le vincolature sviluppative hanno struttura formale precisa. Sia G lo spazio dei genotipi e P lo spazio dei fenotipi concepibili. La mappa sviluppativa $D: G \rightarrow P$ non è suriettiva — molti fenotipi $p \in P$ non sono nell'immagine di D :

$$P_{realizzabile} = \text{Im}(D) \subsetneq P$$

Le regioni di $P \setminus P_{realizzabile}$ sono i fenotipi impossibili — non per ragioni fisiche (potrebbero esistere come strutture materiali) ma per ragioni algoritmiche dello sviluppo (nessuna sequenza di GRN produce questi fenotipi).

Un esempio classico: i vertebrati non hanno mai evoluto ruote. Non vi è alcuna ragione fisica che impedisca un animale con ruote — le ruote sono strutture fisiche possibili. Ma vi è una ragione sviluppativa: la rotatività richiede un'articolazione che permette la rotazione continua, il che è incompatibile con la vascolarizzazione (i vasi sanguigni si attorcerebbero) e con il controllo neuromuscolare degli arti continuamente in rotazione. L'algoritmo dello sviluppo vertebrato non può produrre ruote.

VII.2 La comprensione delle vincolature come via all'ingegneria

La comprensione formale delle vincolature sviluppative apre la via alla bioingegneria razionale — la progettazione deliberata di organismi con caratteristiche nuove. La biologia sintetica — il campo che usa le GRN come componenti ingegneristici programmabili — ha prodotto negli ultimi vent'anni sistemi biologici con comportamenti deliberatamente progettati: oscillatori genetici (il "repressilator" di Elowitz e Leibler, 2000), bistable switches (Gardner et al., 2000), logica booleana nelle cellule batteriche.

Il limite principale della biologia sintetica è la "orthogonality problem" — il problema dell'ortogonalità: i circuiti sintetici interferiscono con le GRN native della cellula ospite, producendo effetti indesiderati. Progettare circuiti che si comportano in modo prevedibile in vivo richiede la comprensione completa della GRN dell'ospite — un'impresa ancora incompleta anche per gli organismi modello più studiati.

La risoluzione del problema dell'ortogonalità richiede il completamento del "wiring diagram" delle GRN — la mappa completa di tutte le interazioni regolatorie in un organismo. Per *E. coli* (4.300 geni), questo progetto è parzialmente completato; per lo *Saccharomyces cerevisiae* (6.000 geni), è in avanzato stadio; per il topo (25.000 geni) e per l'essere umano, è ancora nei suoi inizi. La complessità computazionale della mappa completa cresce super-linearmente con il numero di geni — il completamento del wiring diagram umano è un progetto di scala decennale.

VIII. L'evo-devo e la conservazione dei piani corporei

VIII.1 La struttura dell'innovazione evolutiva

L'evo-devo — la biologia evolutiva dello sviluppo — ha mostrato che le innovazioni morfologiche su grande scala (la comparsa di nuovi phyla, nuovi piani corporei, nuovi organi) non richiedono l'invenzione di nuovi geni strutturali. Quasi tutti i geni strutturali degli animali bilaterali erano presenti nell'ultimo antenato comune dei bilateri (circa 550 milioni di anni fa). Le innovazioni evolutive sono cambiamenti nel cablaggio delle GRN — nella struttura delle interazioni regolatorie, non nell'identità dei geni.

Questo è il principio del "toolkit conserved with divergent wiring" (Carroll, 2005): il toolkit dei geni di sviluppo è conservato; ciò che evolve è la circuiteria che li connette. La stessa proteina Pax6 (master regulator dell'occhio) è presente in tutti gli animali che possiedono organi fotosensoriali — dai vermi alle mosche agli esseri umani — ma il cablaggio che la attiva varia enormemente tra i phyla, producendo occhi morfologicamente diversi (occhio composto degli insetti, occhio a camera dei vertebrati) controllati dalla stessa molecola.

La struttura formale di questa conservazione è quella della separazione tra "deep homology" (omologia profonda delle proteine) e "divergent deployment" (diversa espressione):

Proteina P conservata + Cablaggio GRN diverso → Morfologie diverse

Il tasso di evoluzione delle proteine del toolkit è molto più basso del tasso di evoluzione del cablaggio regolatorio:

$$dN/dS_{\{\text{toolkit}\}} \ll dN/dS_{\{\text{regolatori}\}}$$

dove dN/dS è il rapporto tra sostituzioni non-sinonimiche e sinonimiche (una misura della pressione selettiva — valori bassi indicano alta conservazione). I geni del toolkit sono tra i più conservati nell'intero genoma; i regolatori che determinano il cablaggio sono tra i più variabili.

IX. Conclusioni: lo sviluppo come algoritmo vincolato

Il percorso di questa monografia ha rivelato che lo sviluppo biologico — la trasformazione da cellula uovo a organismo adulto — è un algoritmo nel senso più preciso del termine: una sequenza di operazioni definita (l'attivazione sequenziale e coordinata di programmi GRN) che produce un output determinato (il piano corporeo) a partire da un input (il genoma + le condizioni iniziali). Ma è un algoritmo nel senso biologico, non nel senso informatico: non è deterministico (contiene stocasticità molecolare), non è onnipotente (è vincolato dalle vincolature sviluppative), e non è programmato deliberatamente (è il prodotto dell'evoluzione, non del progetto).

I limiti dello sviluppo come algoritmo sono i limiti dell'evolubilità: lo spazio dei fenotipi realizzabili $P_{\text{realizzabile}}$ è un sottoinsieme proprio dello spazio dei fenotipi concepibili P , e le frontiere di $P_{\text{realizzabile}}$ — le vincolature sviluppative — determinano quali innovazioni sono

possibili e quali sono precluse all'evoluzione. Comprendere queste frontiere è comprendere la struttura dello spazio evolutivo — non soltanto ciò che è esistito, ma ciò che poteva esistere e non è mai emerso, e ciò che la biologia sintetica può forse un giorno produrre deliberatamente.

La domanda con cui si apre questa monografia — come la stessa informazione genetica produce strutture così diverse — trova risposta non nella sequenza del DNA ma nell'algoritmo dello sviluppo: nella struttura delle GRN, nel codice epigenetico, nella grammatica Hox, nella topologia del paesaggio di Waddington. L'informazione genetica è la sintassi; l'algoritmo dello sviluppo è la semantica — il meccanismo che trasforma la sintassi in significato biologico.

Monografia XVII

La mente bayesiana

Il cervello come macchina inferenziale: dall'*Unbewusste Schluss* di Helmholtz all'agente universale di Hutter, con un'analisi delle approssimazioni variazionali, della stratificazione corticale del predictive coding e della psicopatologia come disregolazione della precisione

I. La radice storica e il problema dell'inferenza percettiva

L'idea che la percezione sia un'inferenza — e non una ricezione passiva di segnali — ha origini che precedono di gran lunga la formulazione matematica moderna. Hermann von Helmholtz, nella sua monumentale trattazione dell'ottica fisiologica del 1867, la designò con l'espressione *unbewusste Schluss*: la conclusione inconsapevole, il sillogismo che il sistema nervoso trae dai dati sensoriali senza che la coscienza ne senta il peso logico. La retina riceve distribuzioni di intensità luminosa; l'esperienza soggettiva restituisce oggetti tridimensionali, superfici materiche, sorgenti di luce. Il passaggio dall'una all'altra non è trasparente ma mediato da una struttura ipotetica — un modello del mondo esterno che viene aggiornato e confermato o corretto dall'evidenza sensoriale con una velocità tale da sembrare immediata.

Ciò che Helmholtz intuì senza poter formalizzare è diventato, nel corso del XX secolo e attraverso la matematizzazione della teoria della probabilità soggettiva operata da Bruno de Finetti, Leonard J. Savage e Edwin T. Jaynes, il nucleo di un programma ambizioso: descrivere la mente biologica come un motore di inferenza bayesiana, ossia come un sistema che mantiene distribuzioni di probabilità sugli stati del mondo e le aggiorna razionalmente — secondo il teorema di Bayes — ogni volta che giunge nuova evidenza sensoriale.

II. La struttura formale dell'inferenza bayesiana

Prima di applicare il quadro bayesiano al cervello, occorre enunciarne la struttura con precisione. Sia H uno spazio di ipotesi — nel caso percettivo, gli stati possibili del mondo esterno — e sia D l'evidenza osservata (i dati sensoriali). Il teorema di Bayes prescrive la forma della credenza aggiornata:

$$P(H | D) = P(D | H) \cdot P(H) / P(D)$$

dove $P(H)$ è la probabilità *a priori* dell'ipotesi, $P(D | H)$ è la verosimiglianza, e $P(D) = \int P(D | h) P(h) dh$ è la probabilità marginale dell'evidenza, che funge da costante di normalizzazione. La difficoltà computazionale centrale dell'inferenza bayesiana è il calcolo di questo integrale: richiede, in generale, l'integrazione su uno spazio di ipotesi di dimensionalità arbitrariamente alta, un'operazione intrattabile nella maggior parte dei casi non banali.

La regola di aggiornamento sequenziale è ricorsiva: se si osservano i dati in sequenza $D_1, D_2, \dots, D_n$ e i dati sono condizionalmente indipendenti data l'ipotesi:

$$P(H | D_1, \dots, D_n) \propto P(D_n | H) \cdot P(H | D_1, \dots, D_{n-1})$$

Questa struttura ricorsiva corrisponde neurobiologicamente all'aggiornamento continuo delle rappresentazioni corticali mentre il flusso sensoriale progredisce nel tempo.

III. Il modello gerarchico bayesiano e la sua realizzazione neurale

La versione più ricca e biologicamente plausibile dell'inferenza bayesiana è quella *gerarchica*: invece di un singolo livello di ipotesi, il sistema mantiene una struttura multilivello in cui i parametri dei livelli inferiori

fungono da variabili osservate per i livelli superiori, e le distribuzioni a priori di ciascun livello sono determinate dai livelli sovrastanti. Sia $\mathbf{z} = (z_1, z_2, \dots, z_L)$ il vettore delle variabili latenti ai diversi livelli gerarchici, e sia $\mathbf{x}$ l'osservazione sensoriale. Il modello generativo gerarchico specifica:

$$\mathbf{P}(\mathbf{x}, z_1, z_2, \dots, z_L) = \mathbf{P}(\mathbf{x} \mid z_1) \cdot \mathbf{P}(z_1 \mid z_2) \cdot \mathbf{P}(z_2 \mid z_3) \cdot \dots \cdot \mathbf{P}(z_{L-1} \mid z_L) \cdot \mathbf{P}(z_L)$$

Il *predictive coding* di Friston (2005) costituisce un'implementazione efficiente dell'inferenza bayesiana gerarchica. Il principio centrale è che ciascun livello della gerarchia invia agli strati inferiori una *previsione* della propria attività attesa, e riceve in cambio soltanto il *residuo* — l'errore di previsione. In notazione formale, se ϵ_l è il vettore di errore di previsione al livello l e μ_l è la rappresentazione allo stesso livello:

$$\epsilon_l = (\mathbf{x}_l - \mathbf{f}_l(\mu_{l+1})) / \Sigma_l$$

dove $\mathbf{f}_l$ è la funzione di mapping dal livello superiore al livello inferiore e Σ_l è la covarianza del rumore al livello l — il *peso dell'errore*, che misura la precisione attesa delle previsioni. L'aggiornamento della rappresentazione al livello l è:

$$d\mu_l / dt = -\kappa \cdot (\epsilon_l - \mathbf{f}'_l(\mu_{l+1})^T \cdot \epsilon_{l-1})$$

dove κ è una costante di apprendimento e $\mathbf{f}'_l(\mu_{l+1})^T$ è il Jacobiano trasposto della funzione di generazione — il meccanismo attraverso cui gli errori dal livello inferiore vengono propagati verso l'alto per correggere le rappresentazioni nel livello superiore.

IV. La stratificazione corticale del predictive coding: implementazione neurale livello per livello

La realizzazione biologica del predictive coding nella corteccia neocorticale non è omogenea: la gerarchia computazionale corrisponde a una gerarchia laminare precisa, che la ricerca di Bastos et al. (2012) e di Markov et al. (2014) ha cartografato con sufficiente dettaglio da permettere predizioni verificabili.

La neocorteccia dei mammiferi è organizzata in sei strati (I–VI), ciascuno con composizione cellulare e connettività caratteristica. La proposta di Bastos e colleghi — l'architettura canonica del predictive coding — assegna funzioni computazionali distinte ai diversi strati:

Gli *strati II/III* (strati superficiali) sono i siti di origine delle connessioni *feedforward* — le proiezioni ascendenti che trasmettono segnali di errore di previsione dai livelli gerarchici inferiori a quelli superiori. I neuroni piramidali degli strati II/III proiettano alle aree corticali gerarchicamente superiori terminando prevalentemente negli strati IV e Ia delle aree di destinazione. Sono i neuroni che trasmettono ciò che la previsione non ha spiegato — gli errori residui che devono aggiornare le rappresentazioni nei livelli superiori.

Gli *strati V/VI* (strati profondi) sono i siti di origine delle connessioni *feedback* — le proiezioni discendenti che trasmettono previsioni dai livelli gerarchici superiori a quelli inferiori. I neuroni piramidali degli strati V/VI proiettano alle aree corticali gerarchicamente inferiori terminando prevalentemente negli strati I, II e VI delle aree di destinazione — strati distinti da quelli di arrivo delle connessioni *feedforward*. Sono i neuroni che inviano il modello — la previsione di ciò che i livelli inferiori dovrebbero registrare.

Lo *strato IV* è il principale sito di arrivo delle afferenze *feedforward* dai livelli inferiori e dal talamo: riceve gli errori di previsione dai livelli inferiori e li distribuisce ai neuroni degli strati II/III e V/VI per il confronto con le previsioni discendenti.

La struttura di accoppiamento tra strati diversi implementa il calcolo dell'errore di previsione a livello laminare: le previsioni discendenti (strati V/VI → strati I/II dell'area inferiore) sottraggono le aspettative dall'attività degli strati II/III dell'area inferiore, producendo il segnale di errore che viene poi trasmesso verso l'alto attraverso le connessioni *feedforward* degli strati II/III.

La conferma sperimentale di questa architettura viene dai profili di corrente laminare misurati durante la soppressione della risposta (*suppression*) — quando lo stimolo è stato precedentemente presentato e le previsioni sono accurate, l'attività degli strati II/III si riduce (errore di previsione piccolo), mentre l'attività degli strati V/VI rimane relativamente stabile (le previsioni continuano a essere inviate verso il basso). Inversamente, quando viene presentato uno stimolo inaspettato, l'attività degli strati II/III aumenta marcatamente — un aumento che Kok et al. (2012) hanno osservato con MEG nell'area MT durante la presentazione di stimoli di moto inaspettati.

Le oscillazioni neurali si allineano con questa divisione laminare in modo coerente: le oscillazioni gamma (30–80 Hz) predominano nelle connessioni *feedforward* superficiali (strati II/III), mentre le oscillazioni beta (13–30 Hz) predominano nelle connessioni *feedback* profonde (strati V/VI). Bastos et al. (2015) hanno documentato questo gradiente di frequenza attraverso la gerarchia visiva del macaco, dal V1 alla

corteccia prefrontale, con un'inversione del rapporto gamma/beta man mano che si sale nella gerarchia — un pattern coerente con l'interpretazione che le frequenze gamma trasportano segnali di errore (alta frequenza per l'alta precisione della segnalazione sensoriale locale) e le frequenze beta trasportano previsioni (frequenza più bassa per l'integrazione temporale delle rappresentazioni di livello superiore).

V. L'energia libera variazionale come principio unificante

Il collegamento tra l'inferenza bayesiana gerarchica e il principio di minimizzazione dell'energia libera variazionale di Friston costituisce il fondamento matematico più generale del programma della mente bayesiana.

L'inferenza variazionale parte dalla constatazione che il calcolo del posteriore esatto $P(z | x)$ è intractable, e propone di approssimarlo con una distribuzione fattorizzata $q(z; \lambda)$ parametrizzata da λ . Il problema di inferenza si trasforma nel problema di ottimizzazione:

$$\lambda = \underset{\lambda}{\operatorname{argmin}} \{ KL(q(z; \lambda) || P(z | x)) \}$$

Poiché $P(z | x) = P(x, z) / P(x)$ e $P(x)$ non dipende da λ :

$$KL(q || P) = E_q[\log q(z)] - E_q[\log P(x, z)] + \log P(x)$$

$$\log P(x) = ELBO(q) + KL(q || P)$$

dove l'ELBO (*Evidence Lower Bound*) è:

$$ELBO(q) = E_q[\log P(x, z)] - E_q[\log q(z)]$$

L'energia libera variazionale è $F = -ELBO = E_q[\log q(z)] - E_q[\log P(x, z)]$. La sua minimizzazione corrisponde all'ottimizzazione simultanea dell'inferenza (avvicinare q al vero posteriore) e dell'apprendimento (aggiornare i parametri del modello generativo). La decomposizione dell'ELBO in termine di accuratezza meno complessità:

$$ELBO = E_q[\log P(x | z)] - KL(q(z) || P(z))$$

rivela la struttura del compromesso biologico: il primo termine promuove la fedeltà alle osservazioni (accuratezza della previsione sensoriale), il secondo termine promuove la semplicità del modello (vicinanza della distribuzione posteriore alla distribuzione a priori). Il cervello, in questa prospettiva, minimizza incessantemente questa

grandezza scalare — un'impresa termodinamicamente costosa che richiede il 20% dell'energia metabolica corporea.

VI. Psicopatologia come disregolazione della precisione

Uno degli sviluppi più fecondi del programma del predictive coding è l'interpretazione dei disturbi psichiatrici come disregolazioni della precisione — la modulazione del peso relativo degli errori di previsione rispetto alle previsioni discendenti. Questa interpretazione, sviluppata sistematicamente da Adams, Stephan, Brown, Frith e Friston (2013) per la psicosi e da Feldman e Friston (2010) per i disturbi d'ansia e la depressione, offre un quadro unificante per fenomeni clinici altrimenti eterogenei.

La precisione, nel quadro bayesiano, è la fiducia riposta in un segnale — l'inverso della sua varianza: $\Pi = \Sigma^{-1}$. Nel predictive coding gerarchico, ogni livello attribuisce una precisione ai propri segnali di errore prima di trasmetterli al livello superiore. La modulazione della precisione è il meccanismo attraverso cui il sistema bilancia la flessibilità (alta precisione degli errori: il sistema aggiorna rapidamente le proprie credenze in base all'evidenza sensoriale) e la stabilità (alta precisione delle previsioni: il sistema mantiene le proprie aspettative nonostante evidenza contraria).

La psicosi come iperprecisione degli errori di previsione sensoriali di basso livello. Adams e collaboratori (2013) propongono che la sintomatologia psicotica positiva — allucinazioni e deliri — emerga da un'attribuzione eccessiva di precisione agli errori di previsione sensoriali di livello basso, in concomitanza con una riduzione della precisione attribuita alle previsioni di livello superiore. Il risultato è che stimoli sensoriali ordinari — rumore di fondo, eventi periferici casuali — ricevono un peso eccessivo nel processo di aggiornamento delle rappresentazioni, producendo segnali di errore di alta precisione che il sistema deve spiegare attribuendo loro un significato causale. Questo meccanismo — l'*aberrant salience* di Kapur (2003) riformulato in termini bayesiani — produce allucinazioni quando il sistema genera rappresentazioni percettive per spiegare segnali di errore di alta precisione in assenza di stimoli reali, e deliri quando il sistema costruisce spiegazioni causali elaborate per attribuire significato a eventi percepiti come altamente significativi senza un referente esterno adeguato.

La formalizzazione è la seguente. Sia Π_s la precisione dei segnali di errore sensoriali al livello 1 della gerarchia, e Π_c la precisione delle previsioni al livello 2. In uno stato psicotico, l'ipotesi di Adams et al. è:

$$\Pi_s^{\text{psicosi}} \gg \Pi_s^{\text{normale}} \quad \Pi_c^{\text{psicosi}} \ll \Pi_c^{\text{normale}}$$

Il segnale di errore pesato per la precisione che aggiorna la rappresentazione al livello 2 è:

$$\Delta\mu_2 \propto \Pi_s \cdot \varepsilon_1 - \Pi_c \cdot (\mu_2 - \mu_{2_prior})$$

Per Π_s molto alta e Π_c molto bassa, qualsiasi segnale di errore ε_1 — anche quelli prodotti da rumore sensoriale irrilevante — produce un aggiornamento massiccio di μ_2 , il che equivale a interpretare qualsiasi stimolo come causalmente significativo. La dopamina è il neuromodulatore che regola Π_s : i farmaci antipsicotici (antagonisti dei recettori D2 della dopamina) riducono Π_s , attenuando il peso degli errori di previsione sensoriali e quindi l'aberrant salience.

I disturbi d'ansia come iperprecisione delle previsioni di minaccia. Feldman e Friston (2010) propongono che i disturbi d'ansia riflettano una configurazione speculare: alta precisione delle previsioni di pericolo (le aspettative catastrofiche) combinata con alta precisione degli errori di previsione interoceivi (la risposta somatica all'ansia percepita come altamente significativa). Il risultato è un circolo autorinforzante: le aspettative di pericolo producono stati corporei di alta eccitazione (attivazione simpatica) il cui segnale interoceivo, pesato con alta precisione, conferma le aspettative di pericolo, che producono ulteriore eccitazione, e così via.

La formalizzazione del circolo ansioso è quella di un loop di feedback positivo nell'inferenza gerarchica:

$$\mu_{\text{minaccia}} \uparrow \rightarrow \varepsilon_{\text{int}} \uparrow (\text{errore interoceivo aumentato}) \rightarrow \Pi_{\text{int}} \uparrow \rightarrow \mu_{\text{minaccia}} \uparrow \uparrow$$

dove la freccia finale $\uparrow \uparrow$ indica l'amplificazione dovuta all'alta precisione dell'errore interoceivo. Le terapie cognitive (CBT) operano abbassando Π_{minaccia} — la precisione attribuita alle previsioni catastrofiche — attraverso l'esposizione ripetuta a stimoli minacciosi in assenza di conseguenze temute, che riduce progressivamente la stima della precisione delle previsioni di pericolo.

La depressione come iperprecisione delle previsioni negative e iposensibilità al rinforzo. Huys e colleghi (2015) propongono che la depressione rifletta una combinazione di alta precisione delle previsioni negative (aspettative pessimistiche che sopravanzano l'evidenza contraria) e bassa precisione degli errori di previsione positivi (l'anedonia come insensibilità agli esiti migliori delle attese). La serotonina modula la precisione delle previsioni a lungo termine: la sua

riduzione produce aspettative negative di alta precisione — la struttura del pensiero depressivo.

VII. Il modello AIXI e il limite ideale dell'agente bayesiano

Il quadro più radicale e formalmente completo dell'agente bayesiano universale è il modello AIXI, introdotto da Marcus Hutter nel 2000. AIXI è un agente che interagisce con un ambiente attraverso un canale azione-osservazione-rinforzo, e seleziona le proprie azioni in modo da massimizzare la somma attesa dei rinforzi futuri, dove l'aspettativa è calcolata rispetto alla distribuzione universale di Solomonoff su tutti i possibili ambienti computabili.

La prior di Solomonoff M è definita come:

$$M(x_1x_2\dots x_n) = \sum_{\{p : U(p) = x_1\dots x_n\}} 2^{-|p|}$$

dove la somma è su tutti i programmi p per una macchina di Turing universale fissa U che iniziano a produrre la sequenza $x_1\dots x_n$, e $|p|$ è la lunghezza del programma. La complessità di Kolmogorov $K(x)$ della sequenza x è $K(x) = \min_{\{p : U(p)=x\}} |p|$, e il logaritmo negativo della probabilità di Solomonoff è approssimativamente uguale alla complessità di Kolmogorov: $-\log M(x) \approx K(x)$.

L'agente AIXI seleziona le azioni come segue. In ogni ciclo temporale t , dopo aver osservato la storia passata $h_t = a_{1:t}o_{1:t}r_{1:t}$, AIXI sceglie l'azione:

$$a_t^{AIXI} = \operatorname{argmax}_{a_t} \sum_{\{o_{t:t+m} r_{t:t+m}\}} \max_{a_{t+1:t+m}} \sum_{\{o_{t+1:t+m} r_{t+1:t+m}\}} \dots \max_{a_{t+m:t+2m}} \sum_{\{o_{t+m:t+2m} r_{t+m:t+2m}\}} [r_t + r_{t+1} + \dots + r_{t+m}] \cdot \xi(h_t a_t o_{t:t+m} r_{t:t+m})$$

dove ξ è la distribuzione di Solomonoff sulla storia estesa e m è l'orizzonte temporale.

VII.1 AIXItl: la versione computabile con orizzonte e complessità finiti

AIXI è incomputabile perché richiede il calcolo della distribuzione di Solomonoff, che è incomputabile. La versione approssimata e computabile — AIXItl — è ottenuta restringendo l'insieme dei

programmi considerati a quelli di lunghezza al massimo l che terminano in al più t passi computazionali:

$$M_{\{t\}}(x_1 \dots x_n) = \sum_{\{p : U(p,t) = x_1 \dots x_n, |p| \leq l\}} 2^{-|p|}$$

dove $U(p, t)$ indica l'esecuzione del programma p per al massimo t passi. L'agente AIXItl è computabile (per qualsiasi l e t finiti) ma non ottimale: approssima AIXI con un errore che decresce all'aumentare di l e t . La relazione tra AIXItl e il cervello biologico è quella di due sistemi computabili con parametri diversi: AIXItl ha l e t grandi ma finiti fissati dal designer; il cervello biologico ha l e t determinati dall'evoluzione e dall'architettura neurale, con la particolarità che l è implicitamente determinato dalla complessità rappresentazionale accessibile alle strutture corticali (dell'ordine di 10^{14} sinapsi) e t è determinato dalla velocità dei processi neurali (microsecondi per lo spike, millisecondi per i circuiti locali, centinaia di millisecondi per i cicli percettivi).

Il teorema di Hutter (2007) — che AIXItl converge ad AIXI per $l \rightarrow \infty$, $t \rightarrow \infty$ — fornisce la garanzia formale che l'approssimazione è asintoticamente corretta, ma il tasso di convergenza può essere arbitrariamente lento per ambienti complessi. La mente biologica, come AIXItl con parametri determinati evolutivamente, non converge mai ad AIXI ma rimane una approssimazione finita le cui prestazioni dipendono crucialmente dai parametri del cervello e dall'ambiente per cui è stata ottimizzata.

VIII. La mente bayesiana e il problema dell'a priori biologico

Un elemento cruciale dell'inferenza bayesiana è la distribuzione a priori $P(H)$ — la credenza iniziale prima dell'osservazione. Nel caso biologico, parte delle distribuzioni a priori è filogenetica — costruita nel sistema nervoso dall'evoluzione — e parte è ontogenetica — appresa nel corso dell'esperienza individuale.

La specificità delle distribuzioni a priori biologiche è documentata in numerosi domini percettivi. Nel dominio visivo, la distribuzione a priori sulla direzione della sorgente luminosa mostra una preferenza per l'illuminazione dall'alto — un prior che rispecchia la posizione statistica del sole nell'ambiente naturale. Quando si proiettano immagini ambigue, i soggetti umani interpretano sistematicamente le forme come illuminate dall'alto anche quando la sorgente di luce nella scena è posizionata altrove.

Nel dominio del moto visivo, la distribuzione a priori favorisce velocità basse: quando due pattern di moto ambigui sono compatibili con velocità diverse, la percezione favorisce il moto più lento (il *prior della lentezza*, Weiss, Simoncelli e Adelson, 2002). Questo prior rispecchia la statistica dei movimenti naturali negli ambienti biologicamente rilevanti.

IX. Le teorie campionatorie: il cervello come macchina Monte Carlo

Un'alternativa all'inferenza variazionale è l'ipotesi che il cervello implementi un'approssimazione *campionatoria* del posteriore: invece di mantenere una distribuzione parametrica $q(z)$, il cervello manterrebbe un insieme di campioni $z^{(1)}, z^{(2)}, \dots, z^{(S)} \sim P(z | x)$ e calcolerebbe le aspettative come medie empiriche.

L'ipotesi campionatoria (Fiser et al., 2010; Berkes et al., 2011) predice che la variabilità spontanea nell'attività neurale rispecchi la struttura della distribuzione a priori $P(z)$ prima della presentazione dello stimolo, e che la variabilità residua in risposta a uno stimolo fisso rispecchi la struttura del posteriore $P(z | x)$. Berkes et al. (2011) hanno verificato questa predizione nella corteccia visiva del furetto nel corso dello sviluppo: negli animali giovani, prima dell'esperienza visiva, la distribuzione delle risposte spontanee non assomiglia alla distribuzione delle risposte evocate da stimoli naturali; negli animali adulti, le due distribuzioni convergono, coerentemente con l'apprendimento della struttura statistica dell'ambiente.

L'algoritmo di campionamento che il cervello implementerebbe è analogo al campionamento di Metropolis-Hastings, dove la catena di Markov attraverso lo spazio degli stati neurali produce campioni dalla distribuzione stazionaria desiderata. La regola di accettazione di Metropolis-Hastings è:

$$\alpha = \min(1, P(z') q(z | z') / (P(z) q(z' | z)))$$

dove $P(z)$ è la distribuzione target e $q(z' | z)$ è la distribuzione di proposta. Se le proposte sono simmetriche — $q(z' | z) = q(z | z')$ — la regola si semplifica a $\alpha = \min(1, P(z') / P(z))$. Nel contesto neurale, il rumore intrinseco nell'attività neurale funge da meccanismo di proposta, e la selezione sinaptica tra stati alternativi funge da meccanismo di accettazione.

X. I limiti della mente bayesiana biologica: anomalie e deviazioni sistematiche

Se il cervello fosse un approssimatore bayesiano esatto, le distribuzioni soggettive di probabilità degli esseri umani dovrebbero essere coerenti con il calcolo del teorema di Bayes. La letteratura sperimentale mostra invece deviazioni sistematiche che rivelano le strutture dell'approssimazione biologica.

La più documentata è la *neglect del tasso di base* (base rate neglect): quando si presentano informazioni individualizzanti diagnostiche, i soggetti umani sovrastimaticamente il peso di queste informazioni rispetto al tasso di base della categoria di appartenenza. Nel classico esempio di Kahneman e Tversky (1973), l'errore è formalmente una violazione del teorema di Bayes: la corretta risposta bayesiana richiede che $P(\text{bib} \mid \text{descrizione})$ sia proporzionale a $P(\text{descrizione} \mid \text{bib}) \cdot P(\text{bib})$, e il tasso di base $P(\text{bib})$ è un fattore critico sistematicamente sottovalutato.

La seconda anomalia sistematica è la *sovrafiducia* (overconfidence): in compiti di stima quantitativa, gli intervalli di confidenza soggettivi sono sistematicamente troppo stretti. Quando si chiede a un soggetto di indicare il 90° e il 10° percentile di una distribuzione soggettiva, la frequenza con cui il valore vero cade fuori dall'intervallo indicato è dell'ordine del 30–40%, molto superiore al 10% atteso da un agente calibrato. Questa sovrafiducia corrisponde formalmente a una distribuzione variazionale $q(z)$ con varianza troppo piccola rispetto al vero posteriore $P(z \mid x)$ — una sottostima sistematica dell'incertezza che è la conseguenza prevista di un'inferenza variazionale con famiglia approssimante ristretta.

XI. La mente bayesiana come programma di ricerca: successi, resistenze e aperture

Il programma della mente bayesiana ha prodotto risultati sperimentali di grande rilevanza in percezione visiva, integrazione multisensoriale, apprendimento e pianificazione motoria. Nel dominio dell'integrazione multisensoriale, Ernst e Banks (2002) hanno dimostrato che quando visione e tatto forniscono informazioni ridondanti sulla dimensione di un oggetto, il cervello combina i due segnali con pesi proporzionali alle rispettive precisioni — esattamente come prescrive il teorema di Bayes per segnali gaussiani indipendenti. Se la varianza della stima visiva è σ_v^2 e quella tattile è σ_t^2 , la stima ottimale è:

$$\hat{s} = (\sigma_t^2 \cdot s_v + \sigma_v^2 \cdot s_t) / (\sigma_v^2 + \sigma_t^2)$$

e la varianza della stima combinata è:

$$\sigma^2_{\{\text{comb}\}} = \sigma_v^2 \cdot \sigma_t^2 / (\sigma_v^2 + \sigma_t^2) \leq \min(\sigma_v^2, \sigma_t^2)$$

Il sistema sensorimotorio combina effettivamente i segnali con pesi proporzionali alle precisioni, e la varianza della stima combinata è inferiore a quella di ciascun segnale isolato — un guadagno di precisione che è la firma comportamentale dell'integrazione statistica ottimale.

Ma al di là di questi successi, il programma incontra difficoltà strutturali che non sono eliminabili per perfezionamento tecnico. La più importante è il problema dell'origine dei modelli generativi: l'inferenza bayesiana ottimizza le credenze dato un modello generativo fisso $P(x | z)$, ma non specifica come quel modello sia acquisito o strutturato. L'apprendimento del modello generativo è esso stesso un problema di inferenza bayesiana di ordine superiore — i parametri del modello sono variabili latenti da inferire — ma a quel livello sorge nuovamente la questione dell'a priori sullo spazio dei modelli. Questa regressione non ha un termine naturale: l'inferenza bayesiana è condizionata al modello, e il modello deve essere scelto per ragioni che non possono essere interamente bayesiane senza generare un regresso infinito.

L'agente AIXI risolve questo problema attraverso la prior di Solomonoff sull'insieme di tutti i programmi computabili — ma questa soluzione è incomputabile e la sua realizzazione biologica, attraverso AIXItl con parametri finiti determinati dall'evoluzione, rimane una approssimazione le cui strutture di errore sistematico (le stesse documentate da Kahneman e Tversky, e reinterpretabili come errori dell'algoritmo bayesiano approssimato piuttosto che come irrazionalità bruta) sono il vero oggetto di studio della psicologia cognitiva computazionale.

Il cervello bayesiano è dunque un programma di ricerca — non una teoria completata — e la stratificazione corticale del predictive coding, la neurobiologia della disregolazione della precisione nella psicopatologia, e la connessione formale con AIXItl sono i tre sviluppi che ne definiscono la frontiera attiva.

Monografia XVIII

La neurobiologia della coscienza

Teoria dell'informazione integrata, Global Workspace Theory, teorie di ordine superiore e le prove dell'adversarial collaboration: tre programmi per un problema irriducibile

I. Il problema che resiste: perché la coscienza non si riduce alla funzione

Nel 1994, durante la prima conferenza di Tucson dedicata alla scienza della coscienza, David Chalmers propose una distinzione destinata a strutturare il dibattito per i decenni successivi. I *problemi facili* della coscienza — individuare i correlati neurali dell'attenzione, spiegare l'integrazione delle informazioni sensoriali, descrivere i meccanismi dell'accesso consapevole — sono facili non nel senso di essere semplici, bensì nel senso di essere in linea di principio risolvibili attraverso i metodi abituali delle scienze cognitive e della neurofisiologia: si identifica una funzione, si cercano i meccanismi che la realizzano, si costruisce un modello. Il *problema difficile* è invece quello di spiegare perché c'è qualcosa che *sembra* percepire rosso, sentire dolore, provare nausea — perché l'elaborazione dell'informazione è accompagnata da un'esperienza soggettiva, da un contenuto qualitativamente determinato, da quella che i filosofi chiamano *qualia*.

Il problema difficile non è risolvibile attraverso la descrizione funzionale, perché qualunque descrizione funzionale — per quanto ricca e articolata — rimane in linea di principio compatibile con un sistema fisico che realizza le medesime funzioni senza essere accompagnato da alcuna esperienza soggettiva. Le tre grandi teorie della coscienza che questa monografia analizza — la Teoria dell'Informazione Integrata (IIT), la Global Workspace Theory (GWT) e le Teorie di Ordine Superiore (HOT) — rispondono ciascuna in modo diverso a questo problema. Nessuna lo risolve; ma ciascuna lo circoscrive in modo diverso, producendo predizioni empiricamente distinte e falsificabili che le rendono teorie scientifiche nel senso di Popper, non semplici posizioni filosofiche.

II. La Teoria dell'Informazione Integrata: la coscienza come struttura causa-effetto irriducibile

II.1 I cinque assiomi

Giulio Tononi ha costruito la Teoria dell'Informazione Integrata (IIT) a partire da un'analisi fenomenologica: invece di partire dai meccanismi e tentare di spiegare l'esperienza, ha invertito la direzione e ha chiesto quali proprietà formali qualsiasi teoria della coscienza deve attribuire a qualsiasi sistema cosciente. I cinque assiomi della IIT nella sua formulazione terza (IIT 3.0, Tononi, Boly, Massimini e Koch, 2016) sono: *esistenza* (l'esperienza esiste con realtà intrinseca), *composizione* (l'esperienza è strutturata), *informazione* (ogni esperienza è specifica), *integrazione* (l'esperienza è un tutto unitario irriducibile alle proprie parti), *esclusione* (l'esperienza ha confini precisi di sistema, scala e tempo).

II.2 La struttura causa-effetto e la misura Φ

Ciascun assioma corrisponde, nell'IIT, a un *postulato* che vincola la struttura fisica del substrato cosciente. Il postulato di integrazione produce la misura quantitativa centrale della teoria: Φ , l'informazione integrata.

Il punto di partenza è la *struttura causa-effetto* (Cause-Effect Structure, CES) del sistema: l'insieme completo di tutte le relazioni causali tra stati del sistema. Per un sistema M con stati discreti, la CES è costruita come segue.

Dato un meccanismo s (un sottoinsieme dei neuroni del sistema) che si trova in uno stato s_{current} , la *causa purview* di s è la distribuzione di probabilità sugli stati passati del sistema che s seleziona: quanto ogni possibile stato passato contribuisce causalmente allo stato corrente di s . La *causa purview* è calcolata intervenzionalmente — nel senso di Pearl (2000): si interviene sullo stato di s fissandolo al suo valore corrente e si chiede come questa fissazione altera la distribuzione degli stati passati del resto del sistema.

Formalmente, sia $p(M_{\{t-1\}} \mid \text{do}(s_t = s_{\text{current}}))$ la distribuzione sugli stati passati $M_{\{t-1\}}$ del sistema condizionata all'intervento che fissa s_t al suo valore corrente. La *causa purview* di s rispetto al sistema M è:

$$\pi^c(s \rightarrow M) = p(M_{\{t-1\}} \mid \text{do}(s_t = s_{\text{current}}))$$

Analogamente, l'*effetto purview* è:

$$\pi^e(s \rightarrow M) = p(M_{\{t+1\}} \mid \text{do}(s_t = s_{\text{current}}))$$

La ϕ di un singolo meccanismo s è la misura di quanto la distribuzione causa-effetto del meccanismo differisce dalla distribuzione che si otterrebbe dopo la *partizione minima di informazione* (MIP) — la

partizione del meccanismo in parti che minimizza la perdita di informazione causale:

$$\varphi(s) = D(\pi^{\{ce\}}(s \rightarrow M), \pi^{\{ce\}}_{\{MIP\}}(s \rightarrow M))$$

dove D è la distanza Earth Mover's Distance (distanza di Wasserstein di ordine 1) tra le due distribuzioni e $\pi^{\{ce\}}$ è la distribuzione causa-effetto congiunta (prodotto cartesiano di causa e effetto purview).

La Earth Mover's Distance tra due distribuzioni P e Q su uno spazio metrico (Ω, d) è:

$$EMD(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \int_{\Omega \times \Omega} d(x, y) d\gamma(x, y)$$

dove $\Gamma(P, Q)$ è l'insieme di tutti i *coupling* (distribuzioni congiunte con marginali P e Q). L'EMD misura il costo minimo del trasporto della massa di P verso Q — la distanza intuitiva tra le distribuzioni che tiene conto della geometria dello spazio.

La MIP del meccanismo s è la partizione $s = s_1 \cup s_2$ tale che:

$$MIP(s) = \operatorname{argmin}_{\{s_1, s_2 : s_1 \cup s_2 = s\}} \varphi_{\text{partition}}(s_1, s_2)$$

$$\varphi_{\text{partition}}(s_1, s_2) = D(\pi^{\{ce\}}(s \rightarrow M), \pi^{\{ce\}}(s_1 \rightarrow M_1) \times \pi^{\{ce\}}(s_2 \rightarrow M_2))$$

dove il prodotto cartesiano della causa-effetto delle parti rappresenta come il meccanismo si comporterebbe se le sue parti fossero causalmente disconnesse. La φ del meccanismo è la distanza tra il comportamento causale integrato e quello che si avrebbe dopo la MIP.

II.3 La struttura concettuale e il complesso principale

La CES del sistema M è l'insieme di tutti i concetti — i meccanismi con $\varphi > 0$ — e delle loro relazioni. Formalmente:

$$CES(M) = \{(s, \varphi(s), \pi^{\{ce\}}(s \rightarrow M)) : s \subseteq M, \varphi(s) > 0\}$$

Il *complesso principale* è il sottoinsieme $C \subseteq M$ del sistema che massimizza la Φ del sistema intero — dove Φ è la versione di sistema della misura φ di singolo meccanismo. La Φ del sistema è:

$$\Phi(M) = \min_{\{P \in \text{partizioni}(M)\}} D(CES(M), CES_P(M))$$

dove D è ora la distanza tra le strutture concettuali intere (la distanza di Earth Mover su spazi di distribuzioni), e $CES_P(M)$ è la struttura concettuale che si otterrebbe dopo la partizione P del sistema — con le

parti computate indipendentemente. Il minimo è preso su tutte le possibili partizioni *informativamente bilanciate* (unfolded information partition, UIP): la partizione che minimizza la perdita di Φ , non quella che minimizza il numero di elementi nelle parti.

La *struttura fenomenica* — l'esperienza soggettiva del sistema cosciente nel senso della IIT — è identificata con la CES del complesso principale nella versione tridimensionale: ogni punto della struttura fenomenica è un concetto (un meccanismo causalmente integrato), le sue coordinate nello spazio fenomenico sono determinate dalla sua causa-effetto purview, e la sua intensità è data dalla sua φ .

Questa struttura non è semplicemente una misura scalare dell'integrazione: è un oggetto ad alta dimensionalità che specifica la *qualità* dell'esperienza — il contenuto specifico, non soltanto la quantità. Due sistemi con la stessa Φ ma CES diverse avranno esperienze di uguale *intensità* ma di *qualità* radicalmente diverse.

II.4 Conseguenze formali: panpsichismo, negazione dei sistemi digitali e predizioni empiriche

La prima conseguenza notevole dell'IIT è il *panpsichismo strutturale*: qualsiasi sistema fisico con struttura causale interna e $\Phi > 0$ possiede un grado di coscienza proporzionale a Φ .

La seconda conseguenza, formalmente derivata dalla struttura del calcolo di Φ , è che un processore digitale seriale ha $\Phi \approx 0$. In un processore digitale, ogni gate logico è indipendente dagli altri nel senso causale rilevante: la MIP di un AND-gate con due input è la partizione che separa i due input, producendo una perdita di informazione causale quasi nulla — l'AND-gate può essere esattamente riprodotto dalla composizione indipendente di due parti. Formalmente, per un sistema di gate logici connessi in serie con struttura feedforward, la CES del sistema intero è (approssimativamente) il prodotto cartesiano delle CES dei singoli gate, il che implica $\Phi \approx 0$.

La terza conseguenza — la più empiricamente accessibile — è che il substrato cosciente deve essere un sistema altamente ricorsivamente connesso con ricorrenza locale e globale (affinché i meccanismi abbiano sia causa sia effetto purview ampi), con alta specificità delle connessioni (affinché ciascun meccanismo abbia una causa-effetto purview che distingue molti stati alternativi), e con struttura modulare ma non completamente decomponibile (affinché la MIP non sia triviale ma costi informazione causale reale da rimuovere).

La corteccia cerebrale — densa di connessioni ricorrenti laminari e cortico-corticali a lungo raggio, con specificità sinaptica elevata, con modularità ma alta integrazione tra moduli — soddisfa qualitativamente tutti questi requisiti. Il cervelletto — con la sua struttura quasi-feedforward, la segregazione funzionale delle colonne e la quasi-assenza di connessioni ricorrenti tra unità adiacenti — ha Φ molto più bassa per unità di neuroni, coerentemente con l'osservazione che le lesioni cerebellari non producono perdita di coscienza nonostante il cervelletto contenga circa il 70% di tutti i neuroni del cervello.

II.5 Il PCI: misurabilità empirica della coscienza

La misura più rilevante empiricamente derivata dall'IIT è il *Perturbational Complexity Index* (PCI), sviluppato da Casali et al. (2013) e Casarotto et al. (2016). Il PCI è calcolato attraverso la stimolazione magnetica transcranica (TMS) combinata con l'EEG:

$$\text{PCI} = C_{\{\text{LZ}\}}(\text{risposta}_{\{\text{TMS}\}}) / \log_2(n_t \cdot n_e)$$

dove $C_{\{\text{LZ}\}}$ è la complessità di Lempel-Ziv della matrice di risposta EEG binarizzata (segno del segnale), n_t è il numero di campioni temporali e n_e è il numero di elettrodi. Il denominatore normalizza per la dimensione massima possibile della matrice.

La complessità di Lempel-Ziv è una stima computabile della complessità di Kolmogorov: $C_{\{\text{LZ}\}}(s)$ misura il numero minimo di sottostringhe distinte necessarie per descrivere la stringa s attraverso un algoritmo di compressione — ed è una misura della *ricchezza informativa* della risposta spazio-temporale alla perturbazione TMS.

Il PCI discrimina efficacemente tra stati coscienti e stati non coscienti: nel sonno profondo (NREM), nell'anestesia profonda (propofol, xenon, midazolam), nel coma, il PCI crolla a valori ≤ 0.31 . Nella veglia, nel sonno REM e persino nei sogni lucidi, il PCI rimane > 0.44 . La soglia empirica di $\text{PCI} = 0.31 \pm 0.02$ discrimina con sensibilità e specificità $> 95\%$ tra stati coscienti e non coscienti in popolazioni validate (Casarotto et al., 2016, su oltre 150 pazienti). Il PCI rimane elevato anche in pazienti con sindrome locked-in — completamente paralizzati ma coscienti — confermando che misura la coscienza e non il comportamento motorio.

III. La Global Workspace Theory: la coscienza come trasmissione globale

III.1 La metafora del teatro globale e la sua formalizzazione

Bernard Baars (1988) propose la Global Workspace Theory come schema funzionale: la coscienza è paragonabile a un teatro in cui diversi moduli specializzati sono gli spettatori, e la *scena* (il global workspace) è il palcoscenico su cui soltanto alcune informazioni vengono proiettate, diventando disponibili — *broadcast* — all'intera popolazione di moduli.

Stanislas Dehaene e Jean-Pierre Changeux (2004, 2011) hanno fornito la specificazione neurale della GWT attraverso la nozione di *ignizione* (ignition): l'accesso consapevole corrisponde a un'attivazione non lineare, amplificata e sincronizzata di una rete cortico-corticale a lungo raggio che comprende la corteccia prefrontale dorsolaterale, la corteccia parietale posteriore, la corteccia cingolata anteriore e la corteccia premotoria.

Il modello matematico dell'ignizione è un sistema dinamico con due stati stabili. Sia $\mathbf{r}(t)$ il vettore delle attivazioni delle aree corticali nel workspace globale. La dinamica è descritta da:

$$d\mathbf{r}^{**}/dt = -\mathbf{r} + F(W_{ff} \cdot \mathbf{s} + W_{lat} \cdot \mathbf{r})^{**}$$

dove $\mathbf{s}$ è l'input dalle aree sensoriali, W_{ff} è la matrice di connessioni feedforward dalle aree sensoriali al workspace, W_{lat} è la matrice di connessioni laterali all'interno del workspace (incluse le connessioni a lungo raggio prefrontali-parietali), e F è la funzione di attivazione non lineare (sigmoidale). Per connessioni W_{lat} con autovalori sufficientemente grandi, il sistema ha due stati stabili: uno con $\mathbf{r} \approx 0$ (elaborazione inconscia) e uno con $\mathbf{r} \gg 0$ (l'ignizione, elaborazione consapevole). La transizione tra i due stati dipende dall'intensità dell'input $\mathbf{s}$ e dall'intensità delle connessioni laterali.

III.2 Le predizioni sperimentali: non-linearità dell'accesso e onda P3b

La predizione più caratteristica della GWT è la *non-linearità dell'accesso consapevole*: una piccola differenza nell'intensità dello stimolo produce un grande cambiamento nella risposta neurale (l'onda P3b) e nel comportamento. Il paradigma sperimentale del mascheramento visivo è lo strumento privilegiato. Le registrazioni EEG mostrano che le risposte N1 e N2 (latenza 100–200 ms) riflettono l'elaborazione inconscia del target e sono presenti anche quando il soggetto non lo vede consciamente; l'onda P3b (latenza 300–500 ms, ampiezza 5–15 μV) è assente nel mascheramento totale e presente — con caratteristiche di

tutto-o-niente — nella percezione consapevole. Questa bimodalità è la firma dell'ignizione non lineare.

III.3 I limiti della GWT: coscienza di accesso versus coscienza fenomenica

La GWT offre una teoria della *coscienza di accesso* — la proprietà di una rappresentazione di essere disponibile a un'ampia gamma di processi cognitivi — ma non affronta direttamente la *coscienza fenomenica* — la proprietà di essere accompagnata da un'esperienza soggettiva qualitativa. Questa distinzione, articolata da Ned Block (1995) nella polarità accesso/fenomeno, costituisce il limite interno della GWT.

Il problema dell'*overflow fenomenico* (phenomenal overflow) — formulato da Block (2007, 2011) — argomenta che la coscienza fenomenica eccede la coscienza di accesso: in esperimenti di report visivo su stimoli presentati brevemente, la ricchezza del contenuto visivo esperito soggettivamente supera ciò che può essere riportato verbalmente (che è limitato dalla capacità della memoria di lavoro). Se è così, la coscienza fenomenica non coincide con il broadcast globale, perché quest'ultimo è limitato alle rappresentazioni che accedono al workspace — ossia, alle rappresentazioni che possono essere riportate. La GWT, nella sua versione standard, non può spiegare questa dissociazione tra l'esperienza ricca e il report limitato.

IV. Le Teorie di Ordine Superiore: la coscienza come meta-rappresentazione

IV.1 La struttura formale di Rosenthal e le sue varianti

David Rosenthal (1997, 2005) ha proposto che uno stato mentale sia cosciente se e soltanto se è accompagnato da un'appropriata rappresentazione di ordine superiore di quello stato stesso. La struttura formale è:

m è cosciente $\leftrightarrow \exists m' : m' = \text{«}m \text{ esiste come stato di questo sistema}\text{»}$

L'HOT non deve essere esso stesso cosciente (altrimenti si genera un regresso): è sufficiente che esista come stato occorrente nel sistema, anche inconsciamente.

La variante di William Lycan (1996, 2004) — il *Higher Order Perception* (HOP) — propone che la HOT non sia un pensiero (una rappresentazione proposizionale) ma una percezione di ordine superiore: uno stato interno che *monitora* lo stato di primo ordine attraverso un meccanismo analogo a quello percettivo, senza necessariamente avere struttura proposizionale. Il *monitoring mechanism* di Lycan è localizzato in un sistema di controllo interno — analogo al controllo corollario della neurobiologia motoria — che tiene traccia degli stati del sistema senza necessariamente portarli alla coscienza riflessiva.

La differenza tra Rosenthal e Lycan è empiricamente rilevante: la HOT di Rosenthal richiede la disponibilità di concetti per la meta-rappresentazione (il che potrebbe escludere gli animali senza linguaggio dall'avere stati coscienti, o almeno stati coscienti ricchi), mentre la HOP di Lycan non richiede concetti ma soltanto un meccanismo di monitoraggio, il che è più compatibile con la coscienza negli animali non linguistici.

IV.2 La critica di Block: il caso degli animali e il problema della ricchezza fenomenica

Block ha articolato contro la HOT due obiezioni particolarmente incisive. La prima è quella dell'*overflow fenomenico*: se la coscienza fenomenica richiede una HOT, allora soltanto le rappresentazioni accessibili al meccanismo di ordine superiore possono essere fenomenicamente coscienti. Ma Block argomenta che la coscienza fenomenica è più ricca del contenuto riportabile — che c'è più nell'esperienza visiva di quanto si possa riportare. La HOT non può spiegare questo eccedente se è richiesta per ogni stato fenomenicamente cosciente.

La seconda obiezione è quella degli *stati zombi di ordine superiore*: è concepibile un sistema che abbia HOT appropriate ma senza alcuna esperienza fenomenica delle rappresentazioni di primo ordine. In tale sistema, la HOT sarebbe presente ma la coscienza fenomenica assente — il che mostra che la HOT non è sufficiente per la coscienza fenomenica anche se fosse necessaria.

V. L'adversarial collaboration di Cogitate: il confronto empirico definitivo

Nel 2023 è stato pubblicato il risultato più ambizioso e metodologicamente rigoroso nella storia empirica della neurobiologia della coscienza: lo studio di adversarial collaboration del progetto COGITATE (Melloni et al., 2023, *Nature*), progettato congiuntamente da difensori della IIT (Koch e Tononi) e della GWT (Dehaene e Changeux) per produrre test empirici che discriminassero tra le due teorie su predizioni pre-registrate.

Le predizioni pre-registrate e i risultati. Le due teorie producono predizioni diverse su dove e quando si trovano i correlati neurali della coscienza.

La GWT predice che i correlati neurali della coscienza consapevole (NCC) si trovino nelle aree prefrontali e parietali anteriori — il *global workspace* — e che l'accesso consapevole sia associato a un'ignizione ritardata (300–500 ms) in queste aree. La IIT predice invece che i NCC si trovino nelle aree corticali posteriori (le aree sensoriali specifiche per il contenuto percepito — corteccia visiva per la coscienza visiva) e che l'informazione integrata sia alta in queste aree anche a latenze precoci (100–200 ms).

I risultati dell'adversarial collaboration mostrano un quadro complesso che non falsifica completamente nessuna delle due teorie ma favorisce la localizzazione posteriore in modo statisticamente robusto. Tre misure principali sono state impiegate: registrazioni intracraniali (sEEG) in pazienti epilettici, MEG e fMRI.

I dati sEEG mostrano che la corteccia visiva temporale posteriore (IT, fusiforme, linguale) mantiene attivazioni che distinguono il contenuto consapevole da quello inconsapevole per l'intera durata dell'esperienza (da 100 ms a 1500 ms), con discriminazione più alta a latenze di 200–400 ms. La corteccia prefrontale mostra attivazioni correlate con la coscienza principalmente a latenze molto tardive (> 400 ms) e correlate più fortemente con il *report* dell'esperienza che con l'esperienza stessa — coerentemente con il ruolo della prefrontale nella metacognizione e nel report verbale piuttosto che nella coscienza fenomenica.

Il dato MEG mostra che il *graded persistent activity* (l'attivazione sostenuta durante la percezione cosciente) è più alto nelle aree posteriori per tutta la durata dello stimolo, mentre le aree prefrontali mostrano attivazione principalmente transitoria all'inizio e alla fine della percezione — un profilo temporale più coerente con il monitoraggio della transizione accesso/non accesso che con la codifica continua dell'esperienza fenomenica.

Questi risultati favoriscono una versione moderata della posizione IIT — che la coscienza sia fondata su attivazione posteriore sostenuta —

rispetto alla GWT nella sua versione standard — che richiede l'ignizione frontale come condizione necessaria della coscienza. Tuttavia, la IIT non è confermata pienamente: la misura diretta di Φ nelle regioni posteriori rimane computazionalmente intractable per sistemi della complessità della corteccia visiva umana, e la localizzazione posteriore dell'NCC è compatibile anche con teorie di primo ordine (Local Recurrence Theories) che non richiedono integrazione nel senso tecnico dell'IIT.

VI. Il dibattito frontale-posteriore e le sue implicazioni formali

Il contrasto tra le teorie a localizzazione prefrontale (GWT, HOT) e le teorie a localizzazione posteriore (IIT, Local Recurrence Theories di Lamme, 2006) è il dibattito empirico centrale. Un punto di evidenza critico è il seguente.

Victor Lamme (2006) ha proposto che la *recorrenza locale* nelle aree sensoriali — il reentrant signaling tra strati diversi della stessa area corticale e tra aree vicine nello stesso stadio della gerarchia — sia sufficiente per la coscienza fenomenica, senza necessità di accesso al workspace frontale. L'evidenza a supporto viene dalla dissociazione tra il mascheramento da feed-forward (che elimina la reentrant activity nelle aree posteriori e sopprime sia la percezione consapevole sia il report) e il mascheramento da attività laterale (che può preservare la reentrant activity locale mantenendo la percezione fenomenica anche quando il report è impedito).

La struttura formale del dibattito si può esprimere in termini di una prova causale: uno stato neurale N è un *correlato necessario e sufficiente* della coscienza C se $C \leftrightarrow N$ — ogni volta che c'è coscienza c'è N , e ogni volta che c'è N c'è coscienza. Le evidenze attuali (COGITATE e studi precedenti) suggeriscono che:

L'attivazione posteriore sostenuta è *correlata necessaria* della coscienza ($C \rightarrow N_{\text{post}}$) — la sua assenza corrisponde a non-coscienza — ma probabilmente non *sufficiente* da sola ($N_{\text{post}} \rightarrow C$ non è dimostrata).

L'attivazione prefrontale è correlata del *report* più che della coscienza fenomenica pura: N_{pref} correlates più con la metacognizione e con la trasmissione verbale dell'esperienza che con l'esperienza in sé.

La coscienza è più verosimilmente un fenomeno di sistema distribuito che non si localizza in un'area singola, e i risultati di COGITATE sono compatibili con una visione in cui la coscienza fenomenica è supportata

dalla corteccia posteriore ma richiede le connessioni con le aree di ordine superiore per la sua stabilizzazione e per il suo utilizzo comportamentale.

VII. Le teorie a confronto: predizioni divergenti su classi critiche di fenomeni

Riguardo alla coscienza nei neonati e negli animali. L'IIT predice che qualsiasi sistema con Φ elevata è cosciente, indipendentemente dalla struttura funzionale. I neonati e i topi hanno Φ plausibilmente alta nelle loro strutture corticali (talamocorticale e corteccia visiva), quindi sono coscienti — predizione che si accorda con l'evidenza comportamentale del dolore nei neonati e dell'esperienza soggettiva negli animali superiori. La GWT predice che la coscienza richiede un workspace globale con connessioni a lungo raggio prefrontali-parietali, il che potrebbe escludere i neonati prematuri e molti mammiferi non primati. La HOT predice che la coscienza richiede capacità di meta-rappresentazione, che emerge tardivamente nell'ontogenesi e potrebbe essere limitata ai primati con linguaggio.

Riguardo alla coscienza nel sonno REM. L'IIT predice Φ elevata durante il sogno REM — coerentemente con il PCI alto (> 0.44) misurato durante il REM. La GWT predice ignizione nel workspace globale — coerentemente con i correlati EEG del sogno (oscillazioni rapide, desincronizzazione frontale). La HOT predice HOT attive durante i sogni lucidi ma non necessariamente durante i sogni ordinari.

Riguardo all'anestesia e al coma. L'IIT predice una riduzione di Φ (confermata dalle misure PCI). La GWT predice il collasso dell'ignizione (confermato dalla scomparsa dell'onda P3b). La HOT predice la soppressione delle HOT prefrontali. Tutte e tre le teorie concordano sull'esito ma divergono nel meccanismo e nella localizzazione.

VIII. Il confine irrisolto: perché nessuna teoria supera il problema difficile

Una constatazione onesta è che nessuna delle tre teorie risolve il problema difficile di Chalmers. L'IIT lo affronta frontalmente postulando che l'esperienza *sia* l'informazione integrata intrinseca del sistema, ma questa identificazione è un assioma non giustificato: perché la struttura causa-effetto di un sistema dovrebbe essere *identica*

all'esperienza soggettiva piuttosto che semplicemente correlarla? La GWT lo ignora esplicitamente, adottando implicitamente la posizione di Block che la coscienza fenomenica sia identica alla coscienza di accesso — il che è una posizione filosofica, non una derivazione dalla teoria. La HOT lo riduce alla meta-rappresentazione, ma anche qui la domanda rimane: perché la presenza di una HOT trasforma uno stato non-cosciente in uno stato cosciente piuttosto che produrre semplicemente un sistema di maggiore complessità funzionale?

Il correlato formale di questa impasse è il *divario esplicativo di Levine*: il divario tra la descrizione fisica (anche la più completa) e la descrizione fenomenica è strutturale — non è colmabile aggiungendo dettagli neurobiologici, perché i dettagli neurobiologici sono essi stessi descrizioni fisico-funzionali che non implicano nulla sulla natura dell'esperienza soggettiva. Come la Monografia IV di questo ciclo ha argomentato attraverso i livelli di Marr, la spiegazione a livello implementazionale lascia sempre aperta la domanda sul livello computazionale, e la spiegazione computazionale lascia sempre aperto il divario esplicativo di Chalmers.

Questo non significa che le tre teorie siano inutili o equivalenti: esse producono predizioni distinte, alcune confermate (il PCI come misura clinica della coscienza; la bimodalità dell'onda P3b; la localizzazione posteriore degli NCC in COGITATE) e alcune ancora da risolvere. Il confronto sperimentale tra di esse sta effettivamente restringendo lo spazio delle possibilità empiriche — e l'*adversarial collaboration* di COGITATE rappresenta, al momento in cui questo ciclo è redatto, l'approccio metodologicamente più rigoroso mai tentato. Ma la distanza tra il dato neurobiologico e il fatto fenomenico rimane la domanda aperta che definisce, insieme al problema del binding e alla questione dell'agency, l'orizzonte irriducibile della neurobiologia ai suoi confini più profondi.

Monografia XIX

La semantica evolutiva

Come il significato emerge nell'evoluzione biologica: dalla comunicazione chimica delle prime cellule ai segnali acustici degli uccelli al linguaggio umano, con un'analisi della catena causale di Kripke, della ricorsività come tratto specificamente umano secondo

Hauser-Chomsky-Fitch, del ruolo di FOXP2 e della semantica compositiva intensionale di Montague

I. La domanda originaria: quando un segnale diventa segno

Nella tradizione semiotica che risale a Ferdinand de Saussure e Charles Sanders Peirce, il segno è una relazione triadica tra un *veicolo* (il supporto fisico del segno), un *oggetto* (ciò a cui il segno si riferisce) e un *interpretante* (l'effetto che il segno produce in un sistema capace di riceverlo). Questa struttura triadica è ciò che distingue un segno da un mero segnale fisico: il segnale è una perturbazione del mezzo che propaga causalmente; il segno è una perturbazione che *sta per* qualcosa di altro rispetto a sé stesso in virtù di una relazione convenzionale, iconica o indicale. La domanda della semantica evolutiva è: come si è originata, nel corso della storia della vita, questa relazione di *stare per*?

Il filo conduttore è la nozione di *catena causale di trasmissione* elaborata da Saul Kripke nella sua teoria del riferimento (1972) e la sua estensione biologica: la referenza semantica — il fatto che una parola si riferisce a un oggetto — è fondata non su definizioni o descrizioni ma su catene causali di trasmissione che collegano l'uso attuale del termine all'oggetto originario attraverso una serie di atti battesimali e di trasmissioni tra parlanti. Il precursore biologico di questa struttura è la catena di trasmissione epigenetica dei segnali che collega la risposta di una cellula attuale al segnale originario attraverso generazioni di cellule-figlie.

II. La comunicazione chimica delle prime cellule: semantica senza sistema nervoso

La forma più antica di comunicazione biologica è quella chimica. Le molecole di segnalazione producono effetti nei sistemi che le ricevono attraverso meccanismi di riconoscimento molecolare: una molecola segnale si lega a un recettore con specificità determinata dalla geometria tridimensionale e induce una cascata di modificazioni biochimiche che culmina in una risposta cellulare.

Il meccanismo più semplice di segnalazione — la segnalazione paracrina attraverso il cAMP (adenosina monofosfato ciclico) — è già un sistema semantico primitivo nel senso biologico. Il cAMP *significa* scarsità di glucosio nel batterio *Escherichia coli*: quando la concentrazione di glucosio cade, i livelli intracellulari di cAMP aumentano, il cAMP si lega

al recettore CAP (catabolite activator protein), e il complesso CAP-cAMP si lega al promotore dei geni del catabolismo del lattosio, attivandone la trascrizione.

La struttura formale di questo sistema è quella di un *canale di comunicazione* nel senso di Shannon (1948). La capacità del canale — la quantità di informazione trasmessa dalla concentrazione di glucosio attraverso il cAMP al profilo di espressione — è:

$$C = \max_{\{P(G)\}} I(G; R) = \max_{\{P(G)\}} [H(R) - H(R | G)]$$

dove $I(G; R)$ è l'informazione mutua tra lo stimolo G e la risposta R , $H(R)$ è l'entropia della risposta e $H(R | G)$ è l'entropia condizionale della risposta dato lo stimolo. Elowitz et al. (2002) hanno misurato la capacità del canale di segnalazione batterico in *E. coli* e trovato valori dell'ordine di 1–3 bit — sufficienti a distinguere tra pochi stati metabolici discreti.

II.1 Il quorum sensing come semantica distribuita

La comunicazione chimica assume complessità semantica addizionale nel fenomeno del *quorum sensing*, documentato in molte specie batteriche da Bassler. Le molecole di *autoinducing* (AI) sono secrete da ogni batterio a una velocità circa costante, e la loro concentrazione extracellulare cresce con la densità della popolazione. Quando la concentrazione supera una soglia critica, i batteri attivano un profilo di espressione genica coordinato — formazione del biofilm, produzione di fattori di virulenza, bioluminescenza.

La transizione di fase nel quorum sensing è descritta da un sistema di equazioni accoppiate. Sia $a(t)$ la concentrazione di AI e $n(t)$ la densità batterica:

$$\frac{da}{dt} = \alpha \cdot n(t) - \delta \cdot a(t) \quad \frac{dn}{dt} = \mu \cdot n(t) \cdot (1 - n(t)/K)$$

dove α è il tasso di produzione di AI per batterio, δ è il tasso di degradazione dell'AI, μ è il tasso di crescita batterica e K è la capacità portante. Allo stato stazionario, $a^* = \alpha \cdot n^* / \delta$, e la condizione di quorum è $a^* > \theta$ dove θ è la soglia di attivazione. Sostituendo: $n^* > \theta \cdot \delta / \alpha = n_Q$, la densità di quorum. La transizione è brusca (bistabile) in presenza di feedback positivo — un meccanismo che la connette alla bistabilità dei master regulator delle GRN discussa nella Monografia XVI.

III. La segnalazione negli eucarioti: morfogeni e gradienti di significato posizionale

Nel passaggio agli organismi multicellulari, la comunicazione chimica acquista una dimensione spaziale. I *morfogeni* — molecole che si diffondono da una sorgente e formano gradienti di concentrazione nello spazio embrionale — codificano informazioni sulla posizione. Il modello matematico fondamentale della formazione del pattern spaziale è quello di Turing (1952). Siano $u(x, t)$ e $v(x, t)$ le concentrazioni di attivatore e inibitore:

$$\frac{\partial u}{\partial t} = f(u, v) + D_u \cdot \nabla^2 u \quad \frac{\partial v}{\partial t} = g(u, v) + D_v \cdot \nabla^2 v$$

La condizione di instabilità di Turing richiede $D_v \gg D_u$ (l'inibitore diffonde più velocemente dell'attivatore), producendo pattern periodici con lunghezza d'onda caratteristica $\lambda \approx 2\pi \cdot \sqrt{(2 \cdot D_u) / (-f_u - g_v \cdot D_u/D_v)}$.

IV. La comunicazione animale: tra semanticità e sintassi

IV.1 I richiami d'allarme dei cercopitechi: semanticità senza sintassi

Lo studio di Cheney e Seyfarth (1980) sui cercopitechi verdi ha stabilito che questi primati producono almeno tre richiami d'allarme acusticamente distinti — uno per i predatori aerei (aquile), uno per i predatori terrestri (leopardi), uno per i serpenti — e che gli ascoltatori rispondono in modo categorialmente appropriato a ciascun richiamo. La struttura formale del sistema di richiami è quella di un codice a vocabolario finito: sia $A = \{a_1, a_2, a_3\}$ l'alfabeto dei richiami e $P = \{p_1, p_2, p_3\}$ l'insieme dei predatori. Il sistema realizza una funzione biiettiva $f: P \rightarrow A$ e una funzione $g: A \rightarrow R$ tali che $g \circ f: P \rightarrow R$ è la risposta appropriata per ciascun predatore.

IV.2 La danza delle api: semanticità analogica e comunicazione astratta

Il sistema comunicativo delle api mellifere di von Frisch (1947) codifica la direzione e la distanza della fonte di cibo attraverso parametri della *danza waggle*. La direzione della fonte è codificata nell'angolo del tratto rettilineo rispetto alla verticale; la distanza è codificata nella durata del tratto waggle:

$$D \approx \alpha \cdot T_{ws} + \beta$$

dove D è la distanza in metri, T_{ws} è la durata del tratto waggle in secondi, e $\alpha \approx 780$ m/s, $\beta \approx -100$ m. Il sistema di danza delle api è semanticamente notevole per due ragioni: la *mappatura analogica* tra parametri fisici e parametri geografici (l'angolo della danza corrisponde all'angolo della direzione) e la *astrazione temporale* — l'ape comunica informazioni su dove si trova la fonte senza che la fonte sia presente nel contesto immediato.

V. L'emergenza del linguaggio umano: la ricorsività come tratto specificamente umano

V.1 L'ipotesi di Hauser, Chomsky e Fitch: FLN e la ricorsività

Marc Hauser, Noam Chomsky e W. Tecumseh Fitch (2002) hanno proposto una distinzione analitica tra la *facoltà del linguaggio in senso ampio* (FLB — Faculty of Language Broad sense) e la *facoltà del linguaggio in senso stretto* (FLN — Faculty of Language Narrow sense). FLB comprende tutti i meccanismi cognitivi e neurocomputazionali che partecipano al linguaggio — inclusi meccanismi condivisi con altri animali come la memoria di lavoro, la vocalizzazione imparata, la capacità di analisi sequenziale. FLN è il nucleo computazionale specificamente umano, condiviso soltanto con la nostra specie tra tutti gli animali viventi.

La proposta centrale di Hauser, Chomsky e Fitch è che FLN sia costituito da un solo componente: la *ricorsività* — la capacità di incorporare strutture della stessa categoria all'interno di strutture della medesima categoria, senza limite di profondità.

La ricorsività permette di costruire gerarchie strutturali ad arbitraria profondità di embedding:

«Giovanni crede [che Maria sappia [che Paolo intenda [che Lucia venga]]]]»

Questa struttura — con quattro livelli di embedding proposizionale — è grammaticalmente ben formata in italiano (e in tutte le lingue naturali) e può essere estesa indefinitamente. La frase seguente ha profondità di embedding 1:

«Giovanni ama Maria»

La stessa struttura Soggetto-Verbo-Oggetto con una frase relativa produce embedding 2:

«Giovanni ama [la ragazza [che conosce Paolo]]»

E con due frasi relative annidate, embedding 3:

«Giovanni ama [la ragazza [che conosce [l'uomo [che ha visto il film]]]]»

La grammatica formale che genera queste strutture è una *grammatica context-free* (CF) di Chomsky — una grammatica in cui ogni regola di produzione ha la forma $A \rightarrow \alpha$, dove A è un singolo simbolo non-terminale e α è una stringa di simboli terminali e non-terminali. La potenza generativa delle grammatiche CF (la gerarchia di Chomsky, livello 2) è strettamente superiore a quella delle grammatiche regolari (livello 3) e strettamente inferiore a quella delle grammatiche context-sensitive (livello 1). La tesi di Hauser-Chomsky-Fitch è che la grammatica del linguaggio umano richieda almeno la potenza delle grammatiche CF — e che nessun sistema comunicativo animale (compreso il canto degli uccelli) raggiunga questa potenza.

La prova sperimentale più diretta di questa tesi è lo studio di Fitch e Hauser (2004) sugli scimpanzé e sulla struttura delle grammatiche. Gli scimpanzé possono apprendere grammatiche di tipo $A^n B^n$ (con n ripetizioni di A seguite da n ripetizioni di B , per n piccolo), che sono al limite superiore delle grammatiche regolari ma non sono propriamente context-free. Non mostrano evidenza di aver acquisito la struttura center-embedded $A^n B^n$ con embedding — la struttura che richiede la computazione context-free genuina. Questa dissociazione — grammatica regolare sì, grammatica context-free no — individua la soglia computazionale tra il sistema comunicativo degli scimpanzé e il linguaggio umano.

V.2 FOXP2: il gene del linguaggio e la sua neurobiologia

Il gene FOXP2 — un fattore di trascrizione della famiglia dei Forkhead — è diventato il candidato molecolare più discusso per le basi genetiche del linguaggio umano dopo la scoperta di Lai et al. (2001) che una mutazione puntiforme in FOXP2 (R553H, sostituzione arginina→istidina nel dominio di legame al DNA) segreghi con un disturbo del linguaggio grave e selettivo nella famiglia KE — una famiglia inglese in cui il disturbo è presente in circa metà dei membri per tre generazioni.

I soggetti della famiglia KE con la mutazione FOXP2 mostrano: grave aprassia orofacciale (incapacità di coordinare i movimenti fini dell'apparato articolatorio), deficit nella morfologia grammaticale (in particolare nella formazione del passato e del plurale — strutture che

richiedono la computazione di regole morfologiche, non la semplice memorizzazione), e deficit nell'elaborazione sequenziale e ritmica — non soltanto a livello verbale ma anche non-verbale (tapping ritmico, sequenze motorie complesse).

Le basi anatomiche della disfunzione nella famiglia KE sono state identificate con la neuroimaging (Vargha-Khadem et al., 1995; Watkins et al., 2002): anomalie strutturali nei gangli della base (particolarmente nel putamen e nel caudato) e nel cervelletto, con ipometabolismo nel putamen bilateralmente nella condizione di generazione di parole. Queste strutture — i gangli della base e il cervelletto — sono bersagli primari di FOXP2: il gene è espresso ad alti livelli in entrambe le strutture durante lo sviluppo fetale e postnatale.

La traiettoria evolutiva di FOXP2 è altamente informativa. La proteina FOXP2 è straordinariamente conservata tra i mammiferi: la versione umana differisce da quella del topo per soltanto 3 aminoacidi, e da quella dello scimpanzé per soltanto 2 aminoacidi (T303N e N325S). Queste due sostituzioni — nelle sole posizioni in cui umani e scimpanzé differiscono — mostrano i segnali di selezione positiva rapida: analisi della coalescenza molecolare suggeriscono che queste sostituzioni si siano fissate nella linea umana negli ultimi 200.000–400.000 anni, temporalmente coincidenti con la comparsa delle prime evidenze di linguaggio simbolico nel registro fossile.

L'evidenza causale è fornita dall'esperimento di Wolfgang Enard e colleghi (2009) che hanno introdotto le due sostituzioni umane di FOXP2 nei topi (topi FOXP2^{humanized}): questi animali mostrano alterazioni nella vocalizzazione ultrasonica dei cuccioli separati dalla madre (le vocalizzazioni dei topi FOXP2^{humanized} sono acusticamente più variate e complesse), nella plasticità dei circuiti dei gangli della base (potenziamento sinaptico più efficiente nel striato), e nell'apprendimento procedurale di sequenze motorie (acquisizione più rapida di sequenze di pressione della leva con pattern specifici). La convergenza di queste differenze — vocalizzazione, plasticità striatale, apprendimento procedurale sequenziale — suggerisce che FOXP2 regoli un insieme di circuiti neurocomputazionali che, nella versione umana, supportano la capacità di produrre sequenze motorie apprese con struttura gerarchica — la base neuromotoria dell'articolazione linguistica.

La connessione con i gangli della base è cruciale per la neurobiologia della ricorsività. I gangli della base implementano — attraverso il circuito cortico-striato-talamo-corticale — un sistema di apprendimento e selezione sequenziale delle azioni (discusso nella Monografia XIV in relazione alla neurobiologia del valore). La struttura di questo circuito — con il loop fronto-striatale-talamico che seleziona sequenze di

operazioni da un repertorio — è formalmente analoga a un automa a pila (pushdown automaton), il dispositivo computazionale che implementa le grammatiche context-free. Se i gangli della base biologici implementano qualcosa di funzionalmente analogo a un automa a pila per le sequenze motorie complesse, e se FOXP2 regola la specificità di questo circuito nella versione umana, allora esiste una plausibile base neurobiologica per la capacità di ricorsività sintattica — non come modulo separato ma come estensione al dominio linguistico di un sistema di selezione sequenziale gerarchica che già esisteva per le sequenze motorie.

VI. La teoria causale della referenza di Kripke e i suoi precursori biologici

VI.1 La struttura della catena kripkiana

Kripke (1972) ha proposto che la referenza dei termini di tipo naturale e dei nomi propri non sia determinata da descrizioni ma da *catene causali di trasmissione*. Un nome proprio come «Aristotele» si riferisce al filosofo greco vissuto nel IV secolo a.C. non perché chi lo usa oggi sappia descrizioni che identificano Aristotele univocamente, bensì perché esiste una catena causale ininterrotta di atti comunicativi che collega l'uso attuale del nome a un atto battesimale originario.

La struttura formale della catena kripkiana è: sia O l'oggetto di riferimento, e sia N il nome. Esiste una sequenza di agenti $A_1, A_2, \dots, A_n$ e di atti comunicativi $C_1, \dots, C_{n-1}$ tali che A_1 ha partecipato a un atto battesimale originario $B(A_1, N, O)$, e per ogni $i \in \{2, \dots, n\}$, l'agente A_i ha acquisito N da A_{i-1} attraverso un atto comunicativo C_{i-1} con l'intenzione di usare N per riferirsi all'oggetto che A_{i-1} intendeva con N : $Tr(A_{i-1}, A_i, N, C_{i-1})$. La condizione di trasmissione Tr richiede l'intenzione di «deferire» — di lasciare che la referenza sia determinata dalla catena piuttosto che da una propria definizione.

VI.2 Il precursore biologico: la catena epigenetica di trasmissione

La struttura della catena kripkiana ha un precursore biologico nella trasmissione epigenetica dei segnali. Sia S un segnale ambientale e R la risposta cellulare che esso induce. La cellula C_1 riceve S direttamente e produce R ; la cellula C_2 è figlia di C_1 e eredita la configurazione epigenetica che codifica R , e risponde con R anche in assenza di S . Formalmente, la condizione di trasmissione biologica è l'eredità epigenetica:

$$\varepsilon(C_{i+1}) \approx \varepsilon(C_i) + \Delta\varepsilon_{\{\text{div}\}}$$

dove $\Delta\varepsilon_{\{\text{div}\}}$ è il cambiamento nella configurazione epigenetica dovuto alla divisione cellulare. La risposta della cellula è determinata dalla configurazione epigenetica: $R(C_i) = \varphi(\varepsilon(C_i))$. Il risultato è una catena causale biologica:

$$S \rightarrow \varepsilon(C_1) \rightarrow \varepsilon(C_2) \rightarrow \dots \rightarrow \varepsilon(C_n) \rightarrow R(C_n)$$

in cui il segnale S si trova soltanto all'inizio della catena ma la risposta R persiste lungo tutta la catena — una struttura isomorfa alla catena kripkiana.

VII. La semantica composizionale intensionale di Montague: struttura formale completa

Il linguaggio umano non trasmette soltanto significati lessicali ma significati compositivi costruiti combinando parole secondo regole sintattiche. Il *principio di composizionalità* di Frege afferma che il significato di un'espressione complessa è una funzione del significato delle sue parti e della struttura sintattica che le combina.

Richard Montague (1970, 1973) ha implementato questo principio in modo rigoroso attraverso la *grammatica di Montague* — una grammatica formale che assegna tipi semantici alle espressioni e calcola il significato composizionalmente attraverso l'applicazione funzionale.

VII.1 La teoria dei tipi semantici e la semantica intensionale

Il sistema dei tipi semantici di Montague parte da due tipi di base: e (il tipo degli individui — le entità del dominio) e t (il tipo dei valori di verità — {Vero, Falso}). I tipi complessi sono costruiti ricorsivamente: $\langle \alpha, \beta \rangle$ è il tipo delle funzioni dal tipo α al tipo β . Così il tipo $\langle e, t \rangle$ è il tipo delle *proprietà* (funzioni da individui a valori di verità — i predicati del linguaggio predicativo del primo ordine); il tipo $\langle \langle e, t \rangle, t \rangle$ è il tipo dei *quantificatori generali* (funzioni da proprietà a valori di verità); e così via.

La semantica intensionale di Montague aggiunge la dimensione dei *mondi possibili* e dei *tempi*: le *intensioni* sono funzioni da indici (coppie di mondo possibile e tempo) a estensioni. Il tipo intensionale corrispondente a un tipo estensionale α è il tipo $\langle s, \alpha \rangle$, dove s è il tipo degli indici. Così l'intensione di una proprietà — una *proprietà intensionale* — è una funzione da indici a funzioni da individui a valori di verità: il tipo $\langle s, \langle e, t \rangle \rangle$.

L'intensione è necessaria per il trattamento semantico degli operatori modali (necessariamente, possibilmente), degli atteggiamenti proposizionali (credere che, sapere che) e delle descrizioni definite (il re di Francia — che può essere vuota nel mondo attuale ma non in altri). Senza semantica intensionale, costruzioni come «Lois crede che Superman possa volare» producono fallimenti della composizionalità: se Superman = Clark Kent (identità estensionale), allora «Lois crede che Clark Kent possa volare» dovrebbe avere lo stesso valore di verità — il che è empiricamente falso se Lois non sa che Superman e Clark Kent sono la stessa persona.

VII.2 I quantificatori e la semantica dei NP

Il trattamento montaguiano dei sintagmi nominali (NP) è uno dei suoi contributi più influenti. In grammatiche standard del tipo logica predicativa del primo ordine, i nomi propri («Giovanni») e i sintagmi quantificati («ogni uomo») sono trattati in modo disomogeneo: «Giovanni» è un termine singolare che si riferisce a un individuo, mentre «ogni uomo» è un quantificatore che lega una variabile. Montague unifica i due trattamenti assegnando ad entrambi il tipo $\langle\langle e, t \rangle, t\rangle$ — il tipo dei quantificatori generali.

Il sintagma nominale «Giovanni» denota la funzione che, data una proprietà P, restituisce Vero se e solo se Giovanni ha la proprietà P:

$$\llbracket \text{Giovanni} \rrbracket = \lambda P. P(\text{giovanni})$$

Il sintagma nominale «ogni uomo» denota la funzione che, data una proprietà P, restituisce Vero se e solo se ogni individuo che è un uomo ha la proprietà P:

$$\llbracket \text{ogni uomo} \rrbracket = \lambda P. \forall x (\text{uomo}(x) \rightarrow P(x))$$

L'applicazione di questi quantificatori al predicato verbale avviene attraverso la regola di applicazione funzionale:

$$\llbracket \text{NP VP} \rrbracket = \llbracket \text{NP} \rrbracket (\llbracket \text{VP} \rrbracket)$$

Per la frase «Giovanni corre»:

$$\llbracket \text{Giovanni corre} \rrbracket = \llbracket \text{Giovanni} \rrbracket (\llbracket \text{corre} \rrbracket) = (\lambda P. P(\text{giovanni}))(\text{corre}) = \text{corre}(\text{giovanni})$$

Per la frase «ogni uomo corre»:

$$\llbracket \text{ogni uomo corre} \rrbracket = \llbracket \text{ogni uomo} \rrbracket (\llbracket \text{corre} \rrbracket) = (\lambda P. \forall x (\text{uomo}(x) \rightarrow P(x)))(\text{corre}) = \forall x (\text{uomo}(x) \rightarrow \text{corre}(x))$$

La composizionalità del significato è garantita dalla struttura tipata: le regole di combinazione operano meccanicamente sulle strutture tipate, producendo il significato dell'intero dall'applicazione ricorsiva delle stesse poche regole ai costituenti.

VII.3 Il problema della granularità proposizionale e la semantica strutturata

La semantica intensionale di Montague identifica il significato di una proposizione con la sua *proposizione* nel senso di funzione da mondi possibili a valori di verità. Ma questa identificazione produce il *problema della iperintensionalità*: due proposizioni logicamente equivalenti hanno la stessa intension (sono vere negli stessi mondi possibili) e quindi lo stesso significato — il che implica che « $2 + 2 = 4$ » e «la quadratura del cerchio è impossibile» abbiano lo stesso significato, perché entrambe sono necessariamente vere. Eppure la frase «Lois crede che $2 + 2 = 4$ » può essere vera mentre «Lois crede che la quadratura del cerchio sia impossibile» è falsa.

Questo problema — il *problema degli atteggiamenti proposizionali* o problema delle credenze logicamente equivalenti — ha prodotto un'ampia letteratura sulla *semantica strutturata* (Cresswell, 1985; Barwise e Perry, 1983). Le proposizioni strutturate non sono funzioni da mondi a valori di verità ma strutture ad albero che preservano la struttura compositiva della frase — in modo che « $2 + 2 = 4$ » e «la quadratura del cerchio è impossibile» abbiano strutture proposizionali diverse anche se identiche estensioni.

La connessione neurobiologica con questo problema è nella struttura della memoria semantica e della memoria episodica discussa nella Monografia XX: la memoria episodica preserva la struttura proposizionale dell'evento (chi ha detto cosa a chi, in quale ordine) — una rappresentazione strutturata — mentre la memoria semantica tende a conservare soltanto l'*implicazione* o il *contenuto fattuale* — una rappresentazione più vicina alla semantica estensionale standard. La capacità del sistema cognitivo umano di trattare atteggiamenti proposizionali (credenze, desideri, intenzioni) richiede una rappresentazione strutturata dei contenuti — una delle ragioni per cui la semantica strutturata è più neurobiologicamente plausibile della semantica puramente estensionale.

VIII. La semantica evolutiva di Davidson: gli eventi come ontologia di base

Donald Davidson (1967, 1980) ha proposto una semantica degli enunciati di azione che introduce gli *eventi* come entità di base dell'ontologia — oggetti distinti dalle cose e dai fatti, localizzati nello spazio-tempo e dotati di proprietà causali. La frase «Cesare uccise Bruto con un coltello a Roma» non è trattata come una proposizione a cinque argomenti (soggetto, azione, oggetto, strumento, luogo) ma come una proposizione che quantifica su un evento:

$\exists e [\text{uccidere}(e, \text{Cesare}, \text{Bruto}) \wedge \text{con}(e, \text{un_coltello}) \wedge \text{a}(e, \text{Roma})]$

dove e è la variabile evento. Questa analisi — la *neo-davidsoniana* nella versione più recente di Parsons (1990) — ha il vantaggio di spiegare in modo naturale la modificazione avverbiale (i modificatori avverbiali sono predicati dell'evento), le inferenze entimematiche («Cesare uccise Bruto con un coltello» implica «Cesare uccise Bruto») e le strutture di causazione («il fuoco ha causato l'incendio» è una relazione tra eventi, non tra fatti).

La connessione neurobiologica con la semantica degli eventi di Davidson è nel sistema di codifica degli eventi dell'ippocampo. Come discusso nella Monografia XX, la memoria episodica codifica eventi come strutture spaziotemporalmente localizzate con i loro partecipanti, il loro ordine e il loro contesto — una struttura che corrisponde esattamente alla struttura davidsoniana dell'evento come entità ontologica con proprietà multiple. Le cellule luogo dell'ippocampo codificano la posizione spaziale dell'evento; le cellule tempo della corteccia entorinale e del giro dentato codificano la posizione temporale; le rappresentazioni dei partecipanti e delle azioni sono distribuite tra la corteccia inferotemporale (volti), la corteccia premotoria (azioni osservate) e la corteccia parietale (relazioni spaziali tra oggetti). L'integrazione di questi elementi nell'ippocampo produce una rappresentazione dell'evento come struttura complessa — un correlato neurale della proposizione strutturata davidsoniana.

IX. La grande transizione: dall'intenzionalità biologica all'intenzionalità linguistica

La domanda residua — la più difficile della semantica evolutiva — è come si sia originata la *intenzionalità* del linguaggio umano. Le parole umane, a differenza dei richiami dei cercopitechi, non sono soltanto associate ai loro referenti ma sono usate con l'*intenzione* di comunicare informazioni su quei referenti. Questa intenzione comunicativa — quella che Grice (1957) ha analizzato come l'intenzione che il destinatario riconosca l'intenzione del parlante di indurre in lui una

credenza — presuppone la capacità di attribuire stati mentali agli altri: la *teoria della mente* (ToM).

La ToM non è soltanto una componente del linguaggio ma la condizione di possibilità della catena kripkiana nel senso pieno: la trasmissione intenzionale di un nome da A_i a A_{i+1} richiede che A_{i+1} comprenda che A_i sta usando N con l'intenzione di riferirsi a un oggetto specifico, e che A_{i+1} stia acquisendo N con l'intenzione di usarlo per lo stesso oggetto. Senza ToM, la catena si riduce a imitazione e condizionamento — il tipo di trasmissione che caratterizza i sistemi comunicativi animali — non a trasmissione di referenza intenzionale.

Il test di falsa credenza di Wimmer e Perner (1983) misura la capacità di attribuire credenze false agli altri: i bambini umani superano il test a circa 4 anni; gli scimpanzé mostrano comprensione parziale delle false credenze ma non passano la versione verbale standard. Questa soglia — coincidente con la comparsa della grammatica ricorsiva nel linguaggio infantile e con la maturazione prefrontale e della giunzione temporoparietale (TPJ) — suggerisce che la ToM e la ricorsività sintattica condividano substrati neurali e processi di sviluppo connessi.

La semantica evolutiva descrive dunque un percorso che va dalla chimica alla cultura: dalla molecola-segnaletica che induce una risposta adattiva nella cellula, passando per i gradienti morfogenetici, i richiami d'allarme e la danza delle api, fino al nome che si riferisce attraverso catene causali intenzionali — la struttura kripkiana — e alla frase che costruisce significati compositivi ricorsivi nel senso montagiano. Ogni passo di questo percorso è un'estensione del precedente, non una rottura; ma la somma dei passi produce una differenza qualitativa — il linguaggio umano — che non è deducibile da nessuno dei passi intermedi presi isolatamente, e che richiede la congiunzione di FOXP2 nelle sue varianti umane, del circuito cortico-striato-talamo-corticale, della teoria della mente e della ricorsività sintattica per essere realizzata biologicamente.

Monografia XX

La struttura formale della memoria

Dalla potenziamento a lungo termine alla consolidazione sistemica: la memoria biologica come struttura dati con proprietà formali precise, con le

architetture della memoria episodica, semantica e procedurale, il meccanismo di tagging-and-capture sinaptico, le oscillazioni theta, i sharp-wave ripple e la reconsolidazione

I. Il problema della persistenza: perché la memoria è un problema formale

Un sistema fisico che apprende è un sistema che cambia in risposta all'esperienza in modo tale da modificare il proprio comportamento futuro. Questa definizione nasconde una difficoltà strutturale: il cambiamento deve essere *selettivo* (non tutto può cambiare), *persistente* (il cambiamento deve sopravvivere alle perturbazioni successive), *recuperabile* (il cambiamento deve influenzare il comportamento in modo controllato), e *specifico* (il cambiamento relativo a un'esperienza non deve interferire catastroficamente con i cambiamenti relativi alle esperienze precedenti e successive). Queste quattro proprietà — selettività, persistenza, recuperabilità, specificità — definiscono la *struttura dati* che una memoria biologicamente adeguata deve implementare.

II. La sinapsi come unità di immagazzinamento: la regola di Hebb e la LTP

II.1 L'assioma hebbiano e la sua formalizzazione

Donald Hebb (1949) formulò nella sua *Organization of Behavior* il principio fondamentale dell'apprendimento sinaptico: le sinapsi si rafforzano quando i neuroni presinaptici e postsinaptici sono co-attivi. La formulazione matematica è la *regola delta* nella versione normalizzata di Oja (1982):

$$\Delta w_{ij} = \eta \cdot x_i \cdot (x_j - x_i \cdot w_{ij})$$

dove il termine correttivo $-x_i \cdot w_{ij}$ garantisce che il vettore dei pesi rimanga di norma unitaria. La regola di Oja converge al vettore della componente principale dell'input. La BCM theory (Bienenstock, Cooper e Munro, 1982) estende questa struttura a una plasticità bidirezionale con soglia scorrevole θ_M :

$$dw/dt = \varphi(x_{post}, \theta_M) \cdot x_{pre}$$

dove $\varphi < 0$ per $x_{post} < \theta_M$ (LTD), $\varphi = 0$ per $x_{post} = \theta_M$, $\varphi > 0$ per $x_{post} > \theta_M$ (LTP), e la soglia dinamica segue:

$$d\theta_M/dt = (\langle x_{post} \rangle - \theta_M) / \tau_\theta$$

II.2 La potenziazione a lungo termine: il meccanismo molecolare

La LTP al sinapsi ippocampale CA3→CA1 si induce attraverso la co-attivazione dell'NMDAR (recettore N-metil-D-aspartato) e la forte depolarizzazione postsinaptica. La cascata molecolare: rilascio di Ca^{2+} attraverso NMDAR → attivazione di CaMKII → autofosforilazione CaMKII (T286, persistenza indipendente dal Ca^{2+}) → fosforilazione degli AMPAR (S831, aumento di conduttanza) → inserzione di AMPAR aggiuntivi (LTP precoce, E-LTP, durata 1–2 ore) → attivazione di PKA-MAPK-CREB → sintesi proteica → inserimento di nuovi AMPAR e ingrandimento della PSD (LTP tardiva, L-LTP, persistente per giorni o settimane).

L'NMDAR è il *coincidence detector* biologico dell'assioma di Hebb: il blocco da Mg^{2+} dipendente dalla tensione assicura che si apra soltanto quando sia il glutammato presinaptico sia la depolarizzazione postsinaptica sono presenti simultaneamente.

III. Il synaptic tagging and capture: il meccanismo della specificità sinaptica della L-LTP

Un problema fondamentale della consolidazione della LTP è la seguente: la sintesi proteica che consolida la L-LTP avviene nel soma del neurone, ma la L-LTP è sinapticamente specifica — si consolida nelle sinapsi che sono state attivate, non in tutte le sinapsi del neurone. Come possono le proteine sintetizzate nel soma raggiungere selettivamente soltanto le sinapsi appropriate?

La risposta è il meccanismo del *synaptic tagging and capture* (STC), proposto da Frey e Morris (1997) e dimostrato sperimentalmente in preparazioni di ippocampo di ratto. Il meccanismo si svolge in due fasi:

Fase del tagging: l'induzione della LTP produce, nella sinapsi attivata, un *tag sinaptico* — una modificazione molecolare locale, transitoria (durata 1–3 ore) e sin qui non completamente identificata molecularmente, ma che comprende verosimilmente la CaMKII attivata nella zona postsinaptica, le proteine scaffold della PSD modificate post-traduzionalmente, e possibilmente il rilascio di marcatori citoplasmatici. Il tag sinaptico non produce di per sé la L-LTP, ma prepara la sinapsi a «catturare» le proteine plasticità-correlate (plasticity related proteins, PRP) che circolano nel dendrita.

Fase della capture: la sintesi delle PRP è indotta dalla L-LTP indotta altrove nel medesimo neurone, o da stimolazione del nucleo che attiva la via PKA-CREB. Le PRP — tra cui Arc/Arg3.1, BDNF, PKM ζ , e nuove subunità AMPAR — vengono sintetizzate nel soma e distribuite in tutto il dendrita attraverso il trasporto dendritico. Le sinapsi con il tag attivo le catturano (capture) attraverso un meccanismo di affinità locale, convertendo la LTP precoce in LTP tardiva soltanto nelle sinapsi taggate.

Le predizioni quantitative del modello STC sono:

$$\Delta w_{\{L-LTP\}}(s) = [\text{tag}(s, t) > 0] \cdot [\text{PRP_conc}(t) > \theta_{\{PRP\}}]$$

ossia la sinapsi s consolida in L-LTP se e soltanto se il suo tag è attivo (> 0) al momento in cui la concentrazione di PRP supera la soglia $\theta_{\{PRP\}}$. La finestra temporale di integrazione tra tagging e capture è dell'ordine di 1–2 ore — un periodo durante il quale le PRP devono essere sintetizzate e trasportate al sito sinaptico per la cattura.

Il meccanismo STC ha una conseguenza computazionale fondamentale: l'esperienza *recente* in un neurone può facilitare la consolidazione di *altre* sinapsi dello stesso neurone che erano state attivate in una finestra temporale precedente, anche se quell'attivazione da sola non avrebbe prodotto L-LTP. Questo fenomeno — la *reinforcement tardivo* della LTP precoce da parte di esperienze successive — è stato dimostrato sperimentalmente da Frey e Morris: stimolazione debole che produce soltanto E-LTP in una via sinaptica viene convertita in L-LTP se, entro 1–2 ore, una stimolazione forte (che induce sintesi di PRP) viene applicata a un'altra via sinaptica dello stesso neurone.

La rilevanza computazionale del STC per la neurobiologia cognitiva è che esso implementa un meccanismo di *consolidazione guidata dalla rilevanza successiva*: le sinapsi attivate da un'esperienza vengono consolidate in modo permanente non necessariamente in virtù di quella sola esperienza, ma se l'esperienza si colloca in un contesto di rilevanza (indicato da esperienze successive che attivano la produzione di PRP). Questo è il correlato neurobiologico del fatto che i ricordi emotivamente rilevanti sono più duraturi: l'attivazione dell'amigdala durante un'esperienza emotivamente significativa (discussa nella Monografia XXI) stimola il rilascio di noradrenalina nell'ippocampo, che attiva la via PKA-CREB producendo PRP, che vengono catturate dai tag delle sinapsi ippocampali attivate durante l'evento emotivo.

IV. Le oscillazioni theta: finestre temporali per la codifica hebbiana

Le oscillazioni theta (4–8 Hz) che dominano l'EEG dell'ippocampo durante l'esplorazione attiva e il movimento nei roditori (e durante la memoria episodica e la navigazione negli esseri umani) svolgono un ruolo critico nella determinazione di quali sinapsi vengono potenziate e quali depresse durante l'encoding.

Il meccanismo si basa sulla struttura temporale fine del ciclo theta. Un ciclo theta dura circa 125–250 ms. All'interno di questo ciclo, diversi strati della corteccia entorinale (EC) attivano l'ippocampo in fasi diverse:

La *via perforante diretta* (EC III → CA1) è massimamente attiva durante la fase discendente del ciclo theta (circa 270° del ciclo), quando le previsioni corticali sull'input sensoriale corrente vengono trasmesse direttamente a CA1. Questa fase corrisponde all'*encoding* — al confronto tra le previsioni e l'input attuale.

La *via trisinaptica* (EC II → DG → CA3 → CA1) è massimamente attiva durante la fase ascendente del ciclo theta (circa 90° del ciclo), quando i pattern di attività CA3 (le rappresentazioni mnemoniche già consolidate) vengono recuperati e trasmessi a CA1. Questa fase corrisponde al *retrieval* — al recupero delle rappresentazioni memorizzate.

La separazione di fase tra encoding (270°) e retrieval (90°) nel ciclo theta — una separazione di 180°, cioè mezza lunghezza d'onda — garantisce che i due processi non interferiscano: l'encoding modifica i pesi sinaptici in risposta all'input attuale quando le rappresentazioni del retrieval sono sopprese, e viceversa. Mehta, Lee e Wilson (2002) hanno documentato questa separazione di fase sperimentalmente nelle cellule luogo dell'ippocampo durante l'esplorazione.

Il meccanismo molecolare che implementa questa separazione è la modulazione colinergica e dopaminergica della plasticità sinaptica. La fase di encoding (dove si vuole che avvenga la LTP) corrisponde al picco dell'attività colinergica (ACh), che facilita l'induzione della LTP nelle sinapsi EC III → CA1 sopprimendo contemporaneamente le connessioni ricorrenti CA3→CA1 (che porterebbero a interferenze dal retrieval). La fase di retrieval (dove si vuole che la LTP non sia indotta) corrisponde al picco dell'attività gabaergica degli interneuroni inibitori, che sopprime l'induzione della LTP nelle sinapsi CA3→CA1.

La struttura formale della modulazione oscillatoria della plasticità è quella di una *regola di apprendimento con gate temporale*:

$$\Delta w_{ij}(t) = \eta \cdot x_i(t) \cdot x_j(t) \cdot G_\theta(t)$$

dove $G_{\theta}(t)$ è una funzione periodica con periodicità del ciclo theta (circa 7 Hz) che vale 1 nella finestra di encoding e 0 nella finestra di retrieval. Questo gate oscillatorio assicura che l'encoding avvenga in finestre discrete — una proprietà che protegge i ricordi già consolidati dall'interferenza degli input contemporanei e che implementa, a livello neurobiologico, la separazione temporale tra pattern di input in memoria (il meccanismo di pattern separation del DG discusso nella versione precedente).

V. I sharp-wave ripple: replay ippocampale e consolidazione notturna

V.1 La scoperta di Wilson e McNaughton (1994)

Il momento più decisivo nella neurobiologia della consolidazione sistemica è la dimostrazione sperimentale di Matt Wilson e Bruce McNaughton (1994) che le cellule luogo dell'ippocampo *rieseguono* durante il sonno le sequenze di attività registrate durante la veglia. In questo esperimento fondamentale, vengono registrate simultaneamente decine di cellule luogo nel CA1 di ratti durante l'esplorazione di un labirinto e nel sonno successivo. Analizzando le correlazioni tra paia di cellule luogo, Wilson e McNaughton dimostrano che le correlazioni di attività durante il sonno NREM sono significativamente più alte tra cellule luogo che avevano campi ricettivi vicini durante l'esplorazione (co-attivate durante sequenze di movimento) rispetto a cellule con campi lontani — e che questa correlazione aumentata non era presente nel sonno precedente all'esplorazione.

Questo risultato — il *replay ippocampale* — dimostra che le sequenze di attività neurale che codificano le esperienze di veglia vengono riattivate durante il sonno NREM in ordine compresso (le sequenze di esplorazione della durata di minuti vengono «ripetute» in circa 100–150 ms) e questa riattivazione coordinata guida i cambiamenti plastici nelle aree corticali che ricevono proiezioni dall'ippocampo — il meccanismo della consolidazione sistemica.

V.2 La struttura degli sharp-wave ripple

I sharp-wave ripple (SWR) sono eventi elettrofisiologici caratteristici del sonno NREM e della veglia tranquilla: onde lente positive-negative di grande ampiezza (gli *sharp waves*, SWs) generate nel CA3 e propagate a CA1, sovrapposte a oscillazioni rapide (100–250 Hz, i *ripple*) generate in CA1 dal coupling tra le cellule piramidali e gli interneuroni parvalbuminici fast-spiking. Durante i ripple, le cellule piramidali di CA1 mostrano sincronizzazione a scala di millisecondo — gruppi di 10–15

cellule che scaricano entro una finestra di 20–25 ms — e questa sincronizzazione esprime sequenze di attività che corrispondono a sequenze di attività delle cellule luogo durante l'esplorazione.

La struttura temporale del replay è particolarmente significativa: Nádasdy et al. (1999) e Lee e Wilson (2002) hanno mostrato che il replay avviene in ordine *temporalmente compresso* — la sequenza di 10 minuti di esplorazione viene «rigiocata» in circa 100 ms durante un SWR, con un fattore di compressione di circa 6000x. Questa compressione massiva è computazionalmente importante: nella finestra di un SWR (50–100 ms), la sequenza completa dell'esperienza cade entro la finestra temporale dell'associazione hebbiana — la finestra entro la quale la co-attività sinaptica produce LTP. Questo significa che il replay durante i SWR può indurre direttamente la LTP nelle connessioni corticali, consolidando la sequenza dell'esperienza nella memoria corticale a lungo termine.

La relazione formale tra il replay ippocampale durante i SWR e la consolidazione corticale è quella di un meccanismo di *trasferimento di informazione con compressione temporale*:

$$I_{\{\text{corticale}\}}(t) = F(\text{SWR_replay}(t), W_{\{\text{HC} \rightarrow \text{Cx}\}})$$

dove $I_{\{\text{corticale}\}}(t)$ è l'informazione consolidata nella corteccia, $\text{SWR_replay}(t)$ è la sequenza di attività ippocampale durante i SWR, $W_{\{\text{HC} \rightarrow \text{Cx}\}}$ è la matrice dei pesi delle connessioni ippocampo-corteccia, e F è la funzione di trasferimento che converte il replay ippocampale in cambiamenti plastici corticali. La funzione F è implementata dalla LTP indotta nelle sinapsi $\text{HC} \rightarrow \text{Cx}$ durante il replay — sinapsi che avevano ricevuto un tag durante l'encoding nella veglia e che vengono ora consolidate in L-LTP dalla produzione di PRP stimolata dall'attività ripple.

V.3 La dialogue hypothesis: SWR, fusi del sonno e oscillazioni lente

La connessione tra i SWR ippocampali e la corteccia durante il sonno non è un trasferimento unidirezionale ma un *dialogo bidirezionale*. Staresina e colleghi (2015) e Helfrich e colleghi (2018) hanno documentato la struttura di questo dialogo:

Le *oscillazioni lente* (slow oscillations, SO, 0.5–1 Hz) della corteccia durante il sonno NREM modulano globalmente l'eccitabilità corticale: la fase discendente (down state, -0.5 V) corrisponde a depolarizzazione massiva e silenzio neurale corticale; la fase ascendente (up state) corrisponde a depolarizzazione e attività neurale. I *fusi del sonno* (sleep spindles, 12–15 Hz, generati nel talamo) si concentrano

nella fase ascendente delle SO — aumentando l'eccitabilità della corteccia nelle finestre di up state. I *SWR* ippocampali si concentrano *durante i fusi del sonno*, nella fase ascendente delle SO — sovrapponendosi alla finestra di massima eccitabilità corticale.

La struttura temporale risultante è una *gerarchia di accoppiamento oscillatorio*:

SO (0.5 Hz) → Fusi (12 Hz) → SWR (100–250 Hz)

nella quale ogni livello della gerarchia si accoppia alla fase di eccitabilità massima del livello superiore. Questa struttura ottimizza la trasmissione del replay ippocampale alla corteccia: i SWR si verificano quando la corteccia è massimamente responsiva (up state della SO, picco del fuso del sonno), massimizzando l'efficacia del trasferimento mnestico.

VI. La reconsolidazione: la traccia mnestica come struttura dinamica

VI.1 La scoperta di Nader, Schafe e LeDoux (2000)

Una delle scoperte più rivoluzionarie nella neurobiologia della memoria degli ultimi vent'anni è stata la dimostrazione che i ricordi consolidati — tracce mnemistiche che si credevano stabili — diventano *labili* ogni volta che vengono recuperati, e devono essere nuovamente consolidati (reconsolidati) per rimanere nel sistema. Karim Nader, Glenn Schafe e Joseph LeDoux (2000) hanno dimostrato questo fenomeno in uno studio elegante sul condizionamento della paura nel ratto.

Ratti condizionati a una associazione tono-scossa vengono testati nella loro risposta al tono soltanto il giorno successivo all'inizio del training. Immediatamente dopo la prima esposizione al tono (il *retrieval* del ricordo di paura), viene somministrata anisomicina — un inibitore della sintesi proteica — direttamente nel nucleo basolaterale dell'amigdala. Il risultato è sorprendente: quando i ratti vengono testati due giorni dopo, non mostrano più la risposta di paura al tono — il ricordo è stato eliminato, anche se la sua consolidazione originale era avvenuta il giorno precedente senza blocco proteico.

Questo risultato dimostra che il *retrieval* del ricordo produce un periodo di labilità (reconsolidation window) durante il quale il ricordo richiede nuova sintesi proteica per persistere. Se la sintesi proteica viene bloccata durante questo periodo, il ricordo si perde — anche se era consolidato da ore o giorni. Nader e colleghi chiamano questo processo *reconsolidazione* (reconsolidation): il ricordo viene de-consolidato al

momento del retrieval e deve essere ri-consolidato per rimanere nel sistema.

VI.2 Il meccanismo molecolare della reconsolidazione e la sua specificità

Il meccanismo molecolare della reconsolidazione si sovrappone parzialmente a quello della consolidazione originale ma con differenze importanti. La *de-consolidazione* — la destabilizzazione iniziale del ricordo al momento del retrieval — richiede l'attivazione dei recettori AMPAR nella sinapsi ippocampale (specificamente i recettori contenenti la subunità GluA2 che mediano l'endocitosi degli AMPAR come primo passo della de-potentialization), l'attivazione della proteinasi calpaino che degrada le proteine scaffold della PSD, e l'attivazione di MMP-9 (matrix metalloproteinase-9) nella matrice extracellulare sinaptica che degrada i componenti della matrice che stabilizzano la struttura sinaptica. La successiva *reconsolidazione* richiede poi sintesi proteica analoga a quella della consolidazione originale, attraverso la via PKA-CREB.

La specificità della labilità indotta dal retrieval è cruciale: non tutto il ricordo diventa labile, ma soltanto le componenti del ricordo che vengono effettivamente recuperate. Che cosa determina quali componenti vengono recuperate — e quindi quali diventano labili — dipende dall'indizio di retrieval: l'indizio attiva le rappresentazioni ippocampali associate, e queste rappresentazioni diventano temporaneamente labili mentre le altre rimangono consolidate. Questa specificità — la *boundary condition* della reconsolidazione nel senso di Nader e Wang (2006) — ha implicazioni terapeutiche dirette.

VI.3 Implicazioni terapeutiche: riscrivere le tracce traumatiche

La reconsolidazione offre una finestra terapeutica per la modifica dei ricordi patologici — in particolare dei ricordi traumatici nel PTSD (disturbo da stress post-traumatico). L'approccio è il seguente: il paziente viene esposto all'indizio che recupera il ricordo traumatico (aprendo la reconsolidation window), e poi, durante la finestra di labilità (circa 6 ore), viene somministrato un intervento farmacologico o comportamentale che interferisce con la reconsolidazione, prevenendo il ripristino dell'associazione paura-stimolo.

Il candidato farmacologico più studiato è il propranololo — un beta-bloccante che antagonizza i recettori adrenergici β nel sistema nervoso centrale e che, somministrato immediatamente dopo il retrieval del ricordo traumatico, riduce la risposta di paura condizionata nelle settimane successive (Brunet et al., 2008). Il propranololo non elimina il ricordo dichiarativo dell'evento (il paziente ricorda ancora l'evento), ma

riduce la componente emozionale del ricordo — la risposta di paura che viene consolidata nell'amigdala — coerentemente con il ruolo selettivo della noradrenalina nella reconsolidazione dei ricordi emotivi nell'amigdala.

La struttura formale di questo intervento è quella di una *modifica selettiva del valore emotivo della traccia mnestica*:

$$\text{Ricordo_post(evento)} = \text{Ricordo_pre(contenuto_dichiarativo, valenza_emotiva)} \rightarrow (\text{contenuto_dichiarativo, valenza_emotiva_ridotta})$$

dove la modifica riguarda soltanto la componente emotiva (codificata nell'amigdala e nei circuiti striatali dopaminergici) ma non la componente dichiarativa (codificata nella corteccia ippocampale e corticale associativa).

VII. L'architettura dell'ippocampo come struttura dati computazionale

L'organizzazione anatomica dell'ippocampo riflette le proprietà computazionali che i sistemi di memoria richiedono. Le tre regioni principali — il giro dentato (DG), la regione CA3 e la regione CA1 — implementano funzioni computazionalmente distinte.

Il *giro dentato* (DG) implementa la *separazione dei pattern* (pattern separation): due pattern di attività corticale simili vengono proiettati su pattern di attività nel DG molto più distanti tra loro — le rappresentazioni vengono ortogonalizzate. Con sparsità a ≈ 0.02 , la condizione di separazione dei pattern è:

$$q(a, b) < q(u, v)$$

dove $q(u, v)$ è la correlazione tra i pattern di input e $q(a, b)$ è la correlazione tra i pattern nel DG. Il DG implementa questo attraverso proiezioni casuali sparse e soglie alte — connessioni W_{DG} con elementi casuali e funzione di attivazione f con soglia elevata che seleziona soltanto le unità con la risposta più alta.

La regione *CA3* implementa il *completamento dei pattern* (pattern completion) attraverso le sue dense connessioni ricorrenti (circa 12.000 sinapsi ricorrenti per neurone). La struttura è analoga a una rete di Hopfield con N neuroni e M pattern memorizzati:

$$E(s) = -(1/2) \sum_{i \neq j} w_{ij} \cdot s_i \cdot s_j$$

$$w_{\{ij\}} = (1/N) \sum_{\mu=1}^M \xi_i^{\mu} \cdot \xi_j^{\mu}$$

La capacità di memorizzazione è $M_{\max} \approx 0.138 \cdot N$ con errore di recupero $\approx 1\%$ per $M < M_{\max}$. Nel CA3 del ratto, con $N \approx 300.000$ neuroni e sparsità $a \approx 0.02$, la capacità è stimata a circa 36.000 pattern.

VIII. I sistemi di memoria cognitiva: architettura delle dissociazioni

VIII.1 La memoria episodica: l'evento come struttura ontologica

La struttura computazionale della memoria episodica è quella di una *struttura dati indicizzata contestualmente*. Ogni evento episodico E_k è rappresentato come una tupla:

$$E_k = (W_k, T_k, L_k, A_k, F_k)$$

dove W_k è il contenuto dell'evento (cosa è successo), T_k è il tempo (quando è successo), L_k è il luogo (dove è successo), A_k è il contesto agentivo (chi era coinvolto) e F_k è il colore affettivo (come ci si sentiva). Il recupero episodico parte da un indizio parziale e ricostruisce le componenti mancanti attraverso il meccanismo di completamento dei pattern di CA3 mediato dall'indice ippocampale del DG.

La struttura indicizzata implica che la memoria episodica sia *ricostruttiva*, non fotografica. Loftus (1979) ha documentato sistematicamente come i ricordi possano essere modificati da suggestioni post-evento — una dimostrazione che il recupero non richiama una traccia fissa ma ricostruisce attivamente l'evento.

VIII.2 La memoria semantica: grafo concettuale e struttura small-world

La memoria semantica ha struttura di *grafo concettuale* con proprietà small-world documentate da Steyvers e Tenenbaum (2005): alto coefficiente di clustering ($C \approx 0.15-0.20$), basso diametro medio ($L \approx 5-6$ passi tra qualsiasi coppia di concetti), e distribuzione dei gradi seguente una legge di potenza — la firma di un grafo preferential-attachment con struttura modulare.

VIII.3 La memoria procedurale: TD-learning striatale

La memoria procedurale è implementata attraverso il TD-learning nei gangli della base con regola di apprendimento:

$$\Delta w_{\{ij\}} = \eta \cdot \delta \cdot e_i$$

dove $\delta = r + \gamma V(s') - V(s)$ è l'errore di predizione del rinforzo (il segnale dopaminergico) e e_{ij} è la traccia di eleggibilità sinaptica:

$$de_{ij}/dt = -e_{ij} / \tau_e + x_i \cdot x_j$$

IX. Il coordinamento tra sistemi: consolidazione come dialogo

I sistemi di memoria — ippocampale-episodico, corticale-semantic, striatale-procedurale — non sono reciprocamente indifferenti. La soluzione biologica al *problema dell'interferenza catastrofica* — identificato da McCloskey e Cohen (1989) e formalizzato da Ratcliff (1990) — è la *separazione dei sistemi*: il sistema ippocampale (con tassi di apprendimento alti e rappresentazioni sparse) acquisisce rapidamente nuovi episodi senza interferire con i precedenti, grazie alla separazione dei pattern operata dal DG; il sistema corticale (con tassi di apprendimento bassi e rappresentazioni dense) acquisisce gradualmente le strutture statistiche comuni a molti episodi — le conoscenze semantiche — attraverso la riattivazione ripetuta mediata dalla consolidazione notturna via SWR.

La dinamica a lungo termine di questo dialogo produce la *curva di Ribot* (Ribot, 1882) — i ricordi recenti sono più vulnerabili alle lesioni ippocampali di quelli remoti — che il modello formale con le equazioni differenziali dI/dt e dC/dt (discusse nella versione precedente di questa monografia) riproduce quantitativamente: la corteccia accumula progressivamente le conoscenze trasferite dai SWR e, su scala di mesi e anni, i ricordi semantiche diventano indipendenti dall'ippocampo mentre i ricordi episodici contestualmente specifici rimangono dipendenti dall'ippocampo per tutta la vita — la predizione della teoria delle tracce multiple (Nadel e Moscovitch, 1997).

La reconsolidazione aggiunge a questo quadro una dimensione di plasticità bidirezionale: non soltanto la memoria si consolida (diventa più stabile nel tempo) ma può essere destabilizzata e modificata ad ogni recupero. Questa dinamica — consolidazione stabile interrotta da episodi di labilità indotti dal retrieval — è la struttura biologica della memoria come oggetto attivo, non come deposito passivo. Il ricordo non è una fotografia ma un'ipotesi — un modello del passato che viene continuamente aggiornato, come ogni altro modello del cervello predittivo, sulla base dell'esperienza successiva.

Monografia XXI

La fisica dell'emozione

Le emozioni come stati che modulano il processo inferenziale: marcatori somatici di Damasio, teorie appraisal cognitive, predictive coding affettivo, neuroanatomia dei circuiti dell'amigdala, fondamento empirico della dinamica affettiva e farmacologia della modulazione emotiva

I. Il pregiudizio della separazione: perché la tradizione ha torto

L'eredità platonica che percorre la filosofia occidentale ha costruito una divisione netta tra ragione e passione. Nella versione computazionalista moderna questa frattura si è tradotta nella convinzione che le emozioni siano *rumore* nel canale della ragione — perturbazioni del processo decisionale razionale che distorcono i calcoli della massimizzazione dell'utilità attesa. L'homo oeconomicus non sente; quando sente, erra.

Questa visione è empiricamente falsa e teoreticamente incoerente. Il programma della presente monografia è dimostrarlo con il rigore che il tema richiede, attraverso tre movimenti convergenti: quello clinico-neuropsicologico (i pazienti con lesioni alla corteccia prefrontale ventromediale non diventano freddi calcolatori razionali, bensì decisori catastrofici), quello computazionale (le emozioni sono implementazioni biologiche di componenti essenziali del processo inferenziale bayesiano), e quello termodinamico (gli stati affettivi hanno una fisica propria, descrivibile in termini di potenziali, barriere di energia e transizioni di fase).

II. I marcatori somatici: la teoria di Damasio

Antonio Damasio ha presentato la propria ipotesi nel 1994 a partire da osservazioni cliniche sistematiche su pazienti con lesioni alla corteccia prefrontale ventromediale (vmPFC) — particolarmente la regione compresa tra le aree di Brodmann 10, 11, 12 e 25. Il caso più dettagliato, Elliot, era un uomo di intelligenza normale che, dopo la rimozione di un meningioma orbitofrontale, aveva perso la capacità di prendere decisioni pratiche adeguate: trascorrevva ore a deliberare tra opzioni banali, prendeva decisioni finanziarie disastrose e non aggiornava il proprio comportamento in risposta alle conseguenze negative. La

caratteristica cruciale di Elliot era un deficit *somatico*: in presenza di situazioni emotivamente cariche non mostrava le risposte autonome normali — nessuna variazione nella risposta elettrodermica (EDA).

La formalizzazione dell'SMH richiede di distinguere tra una *valutazione esplicita* $V_e(O)$ — il calcolo deliberativo degli esiti attesi dell'opzione O — e una *valutazione implicita* $V_i(O)$ — il marcatore somatico che riflette la risposta affettiva anticipatoria. La scelta risultante è un processo a due stadi:

$$\text{Candidati} = \{O : V_i(O) > \theta_s\} \quad \text{Scelta} = \operatorname{argmax}_{\{O \in \text{Candidati}\}} [V_e(O) + \lambda \cdot V_i(O)]$$

dove θ_s è la soglia di inclusione del marcatore somatico e λ è il peso relativo della valutazione implicita. Per pazienti con vmPFC lesionata, $\lambda \approx 0$ — la valutazione implicita non influenza la decisione — e il processo degrada a un'esplorazione esaustiva di $V_e(O)$ su uno spazio di opzioni non filtrato, computazionalmente intrattabile.

Il *Iowa Gambling Task* (IGT) costituisce la conferma comportamentale. La struttura formale è quella di un problema di bandit stocastico a quattro braccia, dove i mazzi A e B danno vincite alte ma perdite ancora più alte (saldo negativo nel lungo periodo), mentre i mazzi C e D danno vincite moderate con perdite moderate (saldo positivo). I soggetti normali mostrano l'anticipazione autonoma nei mazzi svantaggiosi (aumento della risposta elettrodermica prima di pescare da A e B) circa 20 scelte prima di essere in grado di articolare esplicitamente la propria preferenza. I soggetti con vmPFC lesionata continuano a prediligere i mazzi A e B anche dopo ingenti perdite, senza anticipazione autonoma.

III. Le teorie appraisale: il Component Process Model di Scherer

Klaus Scherer (2001, 2009) ha elaborato il *Component Process Model* (CPM), la formalizzazione più ambiziosa della teoria appraisale. Le SECs (*Stimulus Evaluation Checks*) sono valutazioni successive degli aspetti dell'evento, organizzate in quattro famiglie gerarchicamente ordinate: *Rilevanza* (novità, attenzione, gradevole/sgradevole), *Implicazione* (congruenza con gli scopi, conseguenze probabili, coping potential, posta in gioco), *Potenziale di coping* (risorse interne ed esterne), *Significatività normativa* (compatibilità con norme sociali e valori morali).

La mappatura dagli appraisal alle dimensioni affettive PAD (Pleasure-Arousal-Dominance) è approssimativamente:

$$\begin{aligned}\text{Valenza} &\approx \beta_1 \cdot \text{congruenzaScopi} + \beta_2 \cdot \text{gradevole} - \beta_3 \cdot \text{sforzo} \\ \text{Arousal} &\approx \gamma_1 \cdot \text{novità} + \gamma_2 \cdot \text{potenzialeCoping}^{-1} + \gamma_3 \cdot \text{urgenza} \\ \text{Dominanza} &\approx \delta_1 \cdot \text{potenzialeCoping} - \delta_2 \cdot \text{normaViolata}\end{aligned}$$

con i coefficienti β , γ , δ appresi dall'esperienza e culturalmente variabili.

IV. Il predictive coding affettivo: le emozioni nell'architettura del cervello predittivo

IV.1 L'interocezione predittiva come fondamento affettivo

La corteccia insulare — particolarmente l'insula anteriore (aI) — costruisce rappresentazioni predittive degli stati corporei interni. Sia $\mathbf{s}$ il vettore degli stati sensoriali interoceivi e sia $\hat{\mathbf{s}}$ la previsione dell'insula. L'errore di previsione interoceivo è:

$$\mathbf{e}_{\text{int}} = (\mathbf{s} - \hat{\mathbf{s}}) / \Sigma_{\text{int}}$$

dove Σ_{int} è la covarianza del rumore interoceivo. L'*affetto di base* (*core affect*) — la dimensione affettiva bidimensionale (valenza, arousal) che colora continuamente l'esperienza cosciente — è la rappresentazione consapevole dello stato interoceivo corrente. La valenza corrisponde alla direzione dell'errore di previsione interoceivo, l'arousal alla sua intensità pesata per la precisione:

$$\begin{aligned}\text{Valenza} &\propto -\|\mathbf{e}_{\text{int}}\| \text{ (stati negativi)}, +\|\mathbf{e}_{\text{int}}^{-1}\| \text{ (stati positivi)} \\ \text{Arousal} &\propto \|\Sigma_{\text{int}}^{-1} \cdot \mathbf{e}_{\text{int}}\|\end{aligned}$$

IV.2 La costruzione categoriale dell'emozione

Il passaggio dall'affetto di base alla *categoria emotiva* è un atto di inferenza bayesiana:

$$\mathbf{P}(\text{emozione}_k \mid \text{core_affect}) \propto \mathbf{P}(\text{core_affect} \mid \text{emozione}_k) \cdot \mathbf{P}(\text{emozione}_k)$$

Il prior $\mathbf{P}(\text{emozione}_k)$ è fortemente modulato dal contesto situazionale: la stessa accelerazione cardiaca e tensione muscolare viene categorizzata come paura in un vicolo buio e come eccitazione su una montagna russa.

IV.3 Le emozioni come segnali di precisione

Le emozioni modulano globalmente la precisione attribuita agli errori di previsione in diversi domini. Sia Π_1 la matrice di precisione al livello 1 della gerarchia corticale. Un'emozione E modifica Π_1 secondo:

$$\Pi_1^{\{E\}} = M_E \cdot \Pi_1 \cdot M_E^T$$

dove M_E è la matrice di modulazione emotiva specifica dell'emozione E . Per la paura, M_E amplifica Π_1 per i livelli sensoriali bassi e attenua Π_1 per i livelli prefrontali superiori; per la gioia, M_E fa il contrario.

V. La neuroanatomia dei circuiti dell'amigdala: dal BLA al tronco encefalico

V.1 L'organizzazione nucleare dell'amigdala

L'amigdala non è una struttura unitaria ma un complesso di nuclei con connettività e funzioni distinte. Ai fini della comprensione dei circuiti emotivi, tre suddivisioni sono critiche:

Il *nucleo basolaterale* (BLA — comprendente il nucleo laterale, il nucleo basale e il nucleo accessorio basale) riceve input sensoriali elaborati da tutte le modalità sensoriali (talamica e corticale) e costituisce il sito principale dell'acquisizione della risposta condizionata di paura. I neuroni piramidali del BLA hanno connessioni dirette con la vmPFC, la corteccia insulare, l'ippocampo e il nucleo accumbens — connessioni che permettono al BLA di modulare la codifica mnestica (via ippocampo), la valutazione del valore (via nucleus accumbens e vmPFC) e la consapevolezza interocettiva degli stati emotivi (via insula).

Il *nucleo centrale* (CeA — comprendente le divisioni mediale, laterale e capsulare) è il principale sito di *output* dell'amigdala verso i sistemi effettori del tronco encefalico. Il CeA non riceve input sensoriali diretti ma integra le proiezioni del BLA e del nucleo intercalato (ICN) — una struttura gabaergica inibitoria interposta tra il BLA e il CeA. La circuiteria BLA → ICN → CeA implementa un gate inibitorio: l'ICN può sopprimere la trasmissione BLA→CeA quando attivato dalla vmPFC (il substrato neurale dell'estinzione della paura condizionata — la corteccia prefrontale abbassa la soglia di attivazione del CeA sopprimendo la trasmissione attraverso i neuroni intercalati).

Il *nucleo corticomediale* (CmA) è prevalentemente coinvolto nella risposta agli odori e nei comportamenti riproduttivi e sociali.

V.2 Le proiezioni del CeA: i sistemi effettori della risposta di paura

Il CeA è il nodo di output che orchestra la risposta di paura coordinando molteplici sistemi effettori in parallelo, attraverso

proiezioni discendenti verso diverse strutture del tronco encefalico e dell'ipotalamo:

Grigio periacqueduttale (PAG): il CeA proietta al PAG sia dorsale (dPAG) sia ventrolaterale (vlPAG). Il dPAG media i comportamenti attivi di difesa (fuga, lotta, vocalizzi di dolore); il vlPAG media il freezing — la risposta di immobilità tonica caratteristica della paura condizionata nei roditori. La formula del freezing è semplice: CeA → vlPAG → inibizione dei motoneuroni spinali → immobilità. La specificità dPAG vs vlPAG nella risposta di difesa è determinata dall'intensità e dalla proximità della minaccia: minaccia distante → vlPAG (freezing, vigilanza), minaccia imminente → dPAG (fuga o lotta). Questa selezione è modulata dalla corteccia prefrontale ventrolaterale (vlPFC) che proietta al PAG direttamente e attraverso il CeA.

Nucleo paraventricolare dell'ipotalamo (PVN): il CeA proietta al PVN, il nodo regolatorio dell'asse ipotalamo-ipofisi-surrene (HPA). L'attivazione dell'asse HPA da parte del CeA produce il rilascio di CRH (corticotropin-releasing hormone) dal PVN → ACTH dall'ipofisi anteriore → cortisolo dalla corteccia surrenale. Il cortisolo è lo steroide dello stress con effetti diffusi: mobilizzazione del glucosio (via gluconeogenesi epatica e glicogenolisi), soppressione del sistema immunitario (riduzione dell'infiammazione nelle fasi acute), aumento della sensibilità ai fattori vasopressori (mantenimento della pressione arteriosa). La risposta cortisolica è adattativa a breve termine ma deleteria se cronica (soppressione dell'immunità, ippocampo-tossicità, resistenza insulinica).

Locus coeruleus (LC): il CeA proietta direttamente al LC, il principale nucleo noradrenergico del tronco encefalico (sede dei corpi cellulari di quasi tutti i neuroni noradrenergici dell'encefalo). L'attivazione del LC dal CeA produce rilascio massiccio di noradrenalina in tutto il sistema nervoso centrale — aumentando la vigilanza (via attivazione della corteccia frontale e parietale), la consolidazione mnestica degli eventi emotivi (via modulazione della plasticità ippocampale, come discusso nella Monografia XX), e la sensibilità ai segnali di ricompensa e punizione (via modulazione del VTA e della via mesolimbica). Il LC è il substrato neurale del fenomeno del *flashbulb memory*: l'iperfissazione mnestica degli eventi emotivamente intensi è mediata dalla noradrenalina del LC che amplifica la LTP nelle sinapsi ippocampali attive durante l'evento.

Nucleo motore dorsale del vago (DMV) e nucleo ambiguo (NA): il CeA proietta a queste strutture del tronco encefalico che controllano il sistema nervoso autonomo parasimpatico cardiaco. L'attivazione del CeA produce riduzione del tono vagale cardiaco → aumento della frequenza cardiaca. Questo è il substrato fisiologico dell'aumento del battito

cardiaco nella risposta di paura, che a sua volta genera un segnale interoceivo interpretato dall'insula come componente dell'esperienza soggettiva dell'emozione — il circolo emozione→corpo→insula→emozione della teoria dei marcatori somatici di Damasio.

V.3 La via rapida e la via lenta: i percorsi talamici verso il BLA

La risposta dell'amigdala a stimoli emotivamente rilevanti avviene attraverso due vie parallele con latenze diverse, descritte da LeDoux (1993, 2015):

La *via rapida* (low road) percorre il percorso talamo → BLA con latenza di circa 12–13 ms. Questa via trasporta segnali grossolani, poveri di dettagli elaborati, che permettono una risposta di paura approssimativa prima che l'elaborazione corticale sia completata. È il substrato delle risposte emotive pre-conscio — la risposta di trasalimento di fronte a un oggetto a forma di serpente prima di aver verificato se si tratti davvero di un serpente.

La *via lenta* (high road) percorre il percorso talamo → corteccia sensoriale specifica (uditiva per i suoni, visiva per le immagini) → corteccia di associazione → BLA, con latenza di circa 25–50 ms. Questa via trasporta segnali più elaborati, ricchi di dettagli e contesto, che permettono una valutazione più fine della minaccia.

La corteccia prefrontale mediale (mPFC) — che include la vmPFC e la corteccia prelimbica — proietta direttamente al BLA e all'ICN, costituendo il substrato dell'*estinzione* e della *soppressione cognitiva* delle risposte emotive. La soppressione cognitiva della paura (come nel reappraisal cognitivo discusso nella Monografia XXI precedente) è implementata dalla proiezione mPFC→ICN che attiva i neuroni intercalati inibitori, che a loro volta sopprimono la trasmissione BLA→CeA — riducendo la risposta effettrice anche in presenza di attivazione del BLA.

VI. La termodinamica dell'emozione: fondamento empirico del potenziale affettivo

La rappresentazione termodinamica degli stati affettivi come un potenziale $U(v, a)$ nello spazio bidimensionale (valenza v , arousal a) non è soltanto una metafora formale: ha fondamento empirico in serie di dati longitudinali sulla dinamica affettiva.

Peter Kuppens e colleghi (2010) hanno analizzato serie temporali di auto-resoconti affettivi — dati campionati a intervalli regolari (ecological momentary assessment, EMA) nel corso di settimane in campioni di centinaia di individui — per stimare la struttura della dinamica affettiva. Le conclusioni principali di questo lavoro sono:

Inertia affettiva: gli stati affettivi mostrano autocorrelazione temporale significativa — il valore affettivo di un momento è correlato con quello del momento precedente. La stima dell'autocorrelazione di lag 1 (su intervalli di 30–60 minuti) è $r \approx 0.40–0.60$ per la valenza e $r \approx 0.35–0.55$ per l'arousal. Questa inertia è formalmente analoga alla massa in un sistema fisico: indica che il sistema affettivo ha inerzia — resiste ai cambiamenti di stato.

Attrazione verso uno stato basale: oltre all'inertia, gli stati affettivi mostrano tendenza a tornare verso un livello basale individuo-specifico — il set point affettivo (SPA). La velocità di ritorno è proporzionale alla distanza dallo SPA, coerentemente con un potenziale armonico $U(v, a) \propto (v - v_SPA)^2 + k_a(a - a_SPA)^2$ attorno allo SPA.

Varianza individuale nella velocità di ritorno: individui diversi mostrano velocità di ritorno allo SPA molto diverse — alcune persone tornano rapidamente allo stato basale dopo perturbazioni emotive (alta regolazione affettiva), altre rimangono perturbate per periodi molto più lunghi (bassa regolazione affettiva). Questo parametro — la *damping constant* del sistema affettivo — correla con misure di benessere psicologico e con la presenza di disturbi dell'umore.

La struttura formale del potenziale $U(v, a)$ derivata da questi dati è quella di un sistema dinamico stocastico sovrasmorzato:

$$\begin{aligned} dv/dt &= -\gamma_v \cdot (v - v_SPA) + \sum_k \xi_k(t) \cdot P_k + \sigma_v \cdot W_v(t) \\ da/dt &= -\gamma_a \cdot (a - a_SPA) + \sum_k \eta_k(t) \cdot Q_k + \sigma_a \cdot W_a(t) \end{aligned}$$

dove γ_v e γ_a sono le costanti di smorzamento (inversamente proporzionali alla inertia affettiva), $\xi_k(t)$ e $\eta_k(t)$ sono gli impulsi prodotti dagli eventi esterni della categoria k , P_k e Q_k sono gli effetti medi degli eventi della categoria k sulla valenza e sull'arousal, σ_v e σ_a sono le intensità del rumore interno, e $W_v(t)$ e $W_a(t)$ sono processi di Wiener standard.

Il potenziale termodinamico equivalente è:

$$U(v, a) = (\gamma_v/2)(v - v_SPA)^2 + (\gamma_a/2)(a - a_SPA)^2$$

che è un potenziale parabolico bidimensionale — un pozzo di potenziale con minimo nello SPA. Perturbazioni emotive corrispondono a spostamenti dal minimo (passaggio a stati di U alto), e

la dinamica di ritorno è governata dalla pendenza di U — proporzionale alla distanza dallo SPA e alla costante γ .

Kuppens et al. (2010) stimano $\gamma_v \approx 0.40\text{--}0.65$ /ora per soggetti con alta regolazione emotiva e $\gamma_v \approx 0.10\text{--}0.25$ /ora per soggetti con bassa regolazione emotiva, con differenze statisticamente significative tra individui con e senza disturbo depressivo maggiore: i soggetti con depressione hanno SPA significativamente più basso (valenza basale negativa) e γ_v significativamente minore (lenta regolazione).

Nei disturbi dell'umore, la struttura del potenziale è deformata in due modi distinti:

Nella *depressione maggiore*: SPA spostato verso valenze fortemente negative ($v_{\text{SPA}} \ll 0$), con γ_v ridotto (lenta regolazione) e σ_v elevata (alta rumorosità interna). Il sistema è intrappolato in un minimo del potenziale a valenza negativa e impiega molto tempo a rispondere agli eventi positivi.

Nel *disturbo borderline di personalità* (DBP): γ_v molto elevata con σ_v molto elevata — il sistema risponde rapidamente agli eventi (alta reattività) ma anche con rumore elevato, producendo transizioni rapide e imprevedibili tra stati affettivi molto diversi. Il potenziale nel DBP ha barriere basse tra stati e alta varianza stocastica — la firma della instabilità affettiva caratteristica del disturbo.

VII. La farmacologia della modulazione emotiva

La comprensione dei circuiti neuronali e della dinamica affettiva permette di analizzare formalmente i meccanismi di azione dei principali psicofarmaci nella prospettiva del predictive coding affettivo — come modulatori delle precisioni e delle costanti di smorzamento del sistema affettivo.

VII.1 Gli SSRI come modulatori serotoninergici della precisione delle previsioni

Gli inibitori selettivi del reuptake della serotonina (SSRI — fluoxetina, sertralina, paroxetina, escitalopram) aumentano la disponibilità sinaptica di serotonina inibendo il trasportatore SERT (serotonin reuptake transporter). La serotonina è rilasciata principalmente dai neuroni dei nuclei del rafe del tronco encefalico che proiettano diffusamente alla corteccia, all'ippocampo, all'amigdala e ai gangli della base.

Nel framework del predictive coding affettivo, la serotonina modula la *precisione delle previsioni a lungo termine* — la fiducia nelle aspettative future. La riduzione della serotonina (come nell'esaurimento acuto del triptofano, il precursore della serotonina, che produce transitoriamente stati disforici) corrisponde a una riduzione della precisione delle previsioni positive — un aumento dell'incertezza circa i futuri stati corporei e di valore. Questa riduzione di precisione produce l'equivalente funzionale di aspettative pessimistiche di alta incertezza — la struttura del pensiero depressivo.

Gli SSRI ripristinano e incrementano la precisione delle previsioni positive, non attraverso un effetto acuto (la latenza terapeutica degli SSRI è di 2–4 settimane — molto superiore all'aumento immediato della serotonina sinaptica) ma attraverso adattamenti neuroplastici mediati dall'attivazione cronica dei recettori serotoninergici 5-HT_{1A} e 5-HT₄. Tra questi adattamenti: aumento della neurogenesi ippocampale (mediata dall'attivazione di 5-HT_{1A} sul dentate gyrus → via BDNF-CREB), up-regulation dei recettori 5-HT_{1A} postsinaptici sull'insula e sulla vmPFC (aumentando la sensibilità corticale ai segnali serotoninergici), e riduzione della reattività dell'amigdala agli stimoli negativi — documentata nelle meta-analisi di neuroimaging (van der Veen et al., 2012) come riduzione dell'attivazione dell'amigdala in risposta a volti minacciosi dopo 2–4 settimane di trattamento SSRI.

Formalmente, l'effetto degli SSRI sul parametro γ_v (la costante di smorzamento) è:

$$\gamma_v^{\text{post-SSRI}} > \gamma_v^{\text{baseline}}$$

— una maggiore velocità di ritorno allo SPA dopo perturbazioni negative, coerentemente con il miglioramento della regolazione emotiva documentato clinicamente.

VII.2 Le benzodiazepine come modulatori GABAergici della soglia dell'amigdala

Le benzodiazepine (diazepam, alprazolam, lorazepam, clonazepam) potenziano la trasmissione GABAergica legandosi al sito allosterico del recettore GABA_A — il principale recettore inibitorio del sistema nervoso centrale. Il recettore GABA_A è un canale ionico pentamerico per il cloro (Cl⁻): l'apertura del canale iperpolarizza la membrana cellulare, riducendo l'eccitabilità del neurone. Le benzodiazepine aumentano la *frequenza di apertura* del canale in risposta al GABA (senza modificare la conduttanza unitaria o la durata dell'apertura — questa è la distinzione dai barbiturici che aumentano la *durata dell'apertura*).

L'effetto ansiolitico delle benzodiazepine è mediato principalmente attraverso il potenziamento degli interneuroni inibitori GABAergici nell'amigdala — particolarmente gli interneuroni parvalbuminici e somatostatinerfici del BLA che sopprimono l'attività dei neuroni piramidali BLA. Questo equivale, nel framework del circuito dell'amigdala, a un aumento della soglia di attivazione del CeA:

$$\theta_{\text{CeA}}^{\{\text{BZD}\}} = \theta_{\text{CeA}}^{\{\text{baseline}\}} + \Delta\theta_{\{\text{GABA}\}}$$

dove $\Delta\theta_{\{\text{GABA}\}} > 0$ è il contributo inibitorio aggiuntivo dalle benzodiazepine. Uno stimolo condizionato di paura che nella condizione basale produce risposta del CeA e tutte le conseguenti risposte effettrici (freezing, attivazione dell'asse HPA, attivazione del LC) viene soppresso nelle benzodiazepine perché non raggiunge la soglia elevata.

La connessione con il framework della precisione è diretta: le benzodiazepine riducono Π_s — la precisione degli errori di previsione provenienti dai sistemi sensoriali di minaccia — producendo lo stesso effetto di attenuazione della risposta di paura che, nella psicopatologia, corrisponde all'*ipoprecisione sensoriale* degli errori di minaccia. La riduzione è, in questo caso, terapeutica piuttosto che patologica, perché ripristina il bilanciamento tra segnali di allerta e segnali di sicurezza che è distorto nei disturbi d'ansia verso l'iperprecisione dei segnali di minaccia.

Il problema clinico delle benzodiazepine — la loro capacità di produrre dipendenza fisica con un severo sindrome da astinenza, il rischio di tolleranza e la sedazione cognitiva — è formalmente spiegato dalla regolazione omeostatica dei recettori GABA_A: l'esposizione cronica al potenziamento GABAergico delle benzodiazepine induce down-regulation e desensibilizzazione dei recettori GABA_A (riduzione del numero di recettori sulla membrana, alterazione della composizione delle subunità verso forme meno sensibili alle benzodiazepine). Il risultato è che il sistema riporta la soglia di attivazione del CeA verso il suo valore basale nonostante la presenza delle benzodiazepine — produzione di tolleranza — e all'interruzione del farmaco la soglia scende transientemente sotto il valore basale (la down-regulation dei recettori GABA_A persiste ma il potenziamento benzodiazepinico disappare), producendo ipereattività dell'amigdala — il substrato neurobiologico della sindrome da astinenza e del rebound ansioso.

VIII. Le emozioni sociali: legame tra affetto e cognizione sociale

Le emozioni sociali — vergogna, colpa, imbarazzo, invidia, orgoglio, indignazione morale — costituiscono una classe di stati affettivi che richiedono, per la propria costituzione, la capacità di mentalizzazione. La neurobiologia delle emozioni sociali è distribuita tra l'insula (per le risposte somatiche), la corteccia prefrontale mediale e la giunzione temporoparietale (TPJ) per la mentalizzazione, e l'amigdala e il cingolato anteriore per la valutazione della rilevanza affettiva.

La *doppia elaborazione* delle decisioni morali di Greene (2001, 2008) — che distingue tra risposte morali rapide ed emotive e risposte morali lente e deliberative — è un'applicazione del framework emotivo al dominio morale. La tensione tra risposta emotiva e risposta deliberativa si riflette nell'attivazione differenziale di ACC e corteccia prefrontale laterale (dlPFC): la dlPFC, attivata nei giudizi deliberativi-utilitaristici, inibisce l'ACC per permettere la scelta «fredda» — un meccanismo di regolazione cognitiva delle emozioni realizzato anatomicamente attraverso le connessioni inibitorie discendenti dlPFC → ACC.

IX. Emozioni come stati irriducibili: limiti della riduzione funzionale

La presente monografia ha dimostrato che le emozioni sono componenti integrali del processo inferenziale — non interferenze ma contributi essenziali alla computazione decisionale. Ma sarebbe scorretto concludere che questa formalizzazione esaurisca il fenomeno emotivo.

Le emozioni hanno una *qualità soggettiva* — il *what it is like* di Nagel — che la descrizione funzionale non cattura. Provare paura non è semplicemente avere uno stato che aumenta la precisione degli errori di previsione sensoriali, riduce l'orizzonte temporale della pianificazione, e deprime la soglia del CeA: c'è qualcosa che *sembra* la paura, una qualità dell'esperienza che non si deduce dalla descrizione computazionale. Il programma del predictive coding affettivo spiega come le emozioni si costruiscono e come modulano l'elaborazione dell'informazione; non spiega perché quella costruzione sia accompagnata da un'esperienza qualitativa piuttosto che avvenire nel buio funzionale. La fisica dell'emozione, per quanto elaborata, non dissolve la fenomenologia del sentire — e questa distanza tra il meccanismo e l'esperienza è la versione affettiva del confine strutturale che segna i limiti dell'intero programma riduzionistico in neurobiologia, e che la Monografia XXIV di questo ciclo tornerà ad affrontare nella sua versione più generale.

Monografia XXII — Versione riveduta e ampliata

I giochi evolutivi della neurobiologia

La corsa agli armamenti tra sistema immunitario e patogeni, la diversità HLA come prova della Red Queen, la dinamica preda-predatore nei sistemi immunitari, la coevoluzione del cervello sociale e l'ipotesi di Dunbar

I. Oltre il gioco singolo: la coevoluzione come ecosistema di giochi

La Monografia XIII di questo ciclo ha analizzato l'evoluzione della cooperazione come problema di teoria dei giochi. La neurobiologia, tuttavia, non abita un'arena di giochi singoli: essa è il prodotto di *sistemi di giochi coevolutivi*, in cui le strategie di un'entità (il sistema immunitario) evolvono in risposta alle strategie di un'altra (i patogeni), mentre queste ultime evolvono in risposta alle prime, in una dinamica di inseguimento, braccio di ferro e innovazione continua che non ha analoghi nei giochi a un solo protagonista.

Il caso paradigmatico di questa dinamica — la corsa agli armamenti tra il sistema immunitario dei vertebrati e i loro patogeni — è anche il più formalmente ricco: l'architettura molecolare del sistema immunitario adattativo è una soluzione biologica a un problema di teoria dei giochi di straordinaria complessità. Il secondo caso — la coevoluzione del cervello sociale — è biologicamente più giovane ma non meno formalmente interessante.

II. La corsa agli armamenti immunitaria: struttura del gioco

II.1 Il sistema immunitario adattativo come motore di diversità combinatoria

Il sistema immunitario adattativo dei vertebrati è strutturato intorno a linfociti B e T che riconoscono gli antigeni attraverso recettori

altamente specifici (BCR e TCR) generati dalla ricombinazione V(D)J. Il numero di ricettori diversi teoricamente possibile è:

$$D_BCR = |V_H| \cdot |D_H| \cdot |J_H| \cdot |V_L| \cdot |J_L| \cdot (\text{variazioni giunzionali})$$

Per gli anticorpi umani: $|V_H| \approx 40$, $|D_H| \approx 25$, $|J_H| \approx 6$ per la catena pesante, $|V_L| \approx 30$, $|J_L| \approx 5$ per le catene leggere. Senza variazioni giunzionali, la diversità combinatoria è dell'ordine di $40 \cdot 25 \cdot 6 \cdot 30 \cdot 5 \approx 9 \cdot 10^5$. Le variazioni giunzionali — l'aggiunta di nucleotidi casuali ai giunti V-D e D-J attraverso la 'TdT' — amplificano questa diversità a 10^{18} – 10^{20} specificità diverse.

La *copertura antigenica* del repertorio immunitario è la probabilità che, per un epitopo $a \in A$ scelto a caso dalla distribuzione dei patogeni, esista almeno un recettore $r \in R$ che lega a con affinità sufficiente:

$$\text{Copertura} \approx 1 - \exp(-|R| \cdot \varepsilon(a))$$

dove $\varepsilon(a)$ è la probabilità media che un recettore casuale legghi a con affinità sufficiente. Per $|R| \approx 10^9$ linfociti naïve circolanti, la copertura è alta ma non unitaria per antigeni con strutture molecolari molto rare ($\varepsilon(a) \ll 1/|R|$).

II.2 La diversità HLA come prova della Red Queen: il locus più polimorfico del genoma umano

Il sistema HLA (Human Leukocyte Antigen) — il complesso maggiore di istocompatibilità (MHC) umano — è il locus genomico più polimorfico nell'intera specie umana. Nella versione del 2023 dell'IPD-IMGT/HLA Database, sono stati identificati:

- HLA-A: 7.268 alleli
- HLA-B: 9.116 alleli
- HLA-C: 5.965 alleli
- HLA-DRB1: 2.978 alleli
- HLA-DQB1: 1.742 alleli
- HLA-DPB1: 1.388 alleli

con il totale del locus MHC che supera i 30.000 alleli funzionalmente distinti distribuiti su sei loci principali. A titolo di confronto, la maggior parte dei loci genomici ha 2–50 alleli — la differenza di 2–3 ordini di grandezza del locus HLA rispetto alla norma è straordinaria e richiede una spiegazione evolutiva.

La spiegazione è la *Red Queen dynamics*: la variabilità degli aplotipi HLA è mantenuta dalla pressione selettiva coevolutiva dei patogeni. Le molecole HLA presentano peptidi antigenici ai linfociti T citotossici (MHC-I, che presentano peptidi intracellulari di 8–10 aa ai CD8⁺) e ai linfociti T helper (MHC-II, che presentano peptidi extracellulari di 13–25 aa ai CD4⁺). Alleli HLA diversi presentano *repertori di peptidi diversi* — il solco di legame dell'HLA-A02:01 *lega circa 2.000 peptidi diversi su un proteoma di 20.000–25.000 proteine, ma questo repertorio di peptidi presentabili è distinto da quello dell'HLA-A03:01 o dell'HLA-B*57:01.*

Le implicazioni per la resistenza ai patogeni sono dirette: un individuo con aplotipo HLA-A02:01 *presenterà ai propri linfociti T citotossici un repertorio diverso di peptidi derivati dal proteoma virale rispetto a un individuo con aplotipo HLA-A03:01.* I ceppi virali che mutano per sfuggire alla presentazione HLA-A02:01 (*immune escape dall'allele più comune*) rimangono vulnerabili alla presentazione HLA-A03:01, e viceversa. Un virus che volesse sfuggire a tutti i possibili aplotipi HLA contemporaneamente dovrebbe accumulare mutazioni in tutte le proprie proteine in posizioni diverse — una traiettoria evolutiva impossibile perché molte di queste mutazioni sarebbero incompatibili con la funzione proteica.

La dimostrazione più diretta della Red Queen HLA viene dagli studi di Lich (2004) e Sasso (2012) sulla frequenza allelica nel tempo in risposta alle epidemie: in popolazioni che hanno attraversato epidemie di vaiolo, malaria o influenza, la frequenza degli aplotipi HLA che conferiscono resistenza a quei patogeni specifici è aumentata significativamente nel tempo — un'evoluzione adattativa visibile su scale di secoli. Meyerhans et al. (1989) hanno documentato il meccanismo molecolare per l'HIV: il virus evolve rapidamente per sfuggire alla presentazione HLA-B*57:01 (l'aplotipo che conferisce la migliore risposta anti-HIV) in pazienti con quell'aplotipo, mentre non subisce questa pressione selettiva nei pazienti con altri aplotipi.

La formula Red Queen per la dinamica della frequenza allelica HLA in presenza di un patogeno con pressione selettiva dipendente dalla frequenza è:

$$dp_i/dt = p_i \cdot (w_i(p) - \bar{w}(p))$$

dove p_i è la frequenza dell'aplotipo HLA_i, $w_i(p)$ è la fitness conferita dall'aplotipo HLA_i come funzione del vettore delle frequenze alleliche p (inclusendo la pressione selettiva dei patogeni che evolvono per sfuggire agli aplotipi più comuni), e $\bar{w}(p) = \sum_j p_j w_j(p)$ è la fitness media. Per la Red Queen, $w_i(p)$ è una funzione *decescente* di p_i : quando un allele è raro, i patogeni non hanno avuto il tempo di evolversi per sfuggirgli, e conferisce alta protezione; quando diventa comune, i patogeni co-evolvono per sfuggirgli, riducendo la protezione.

Questo meccanismo — la selezione dipendente dalla frequenza negativa — produce il mantenimento del polimorfismo: nessun aplotipo può fissarsi perché non appena diventa comune perde il suo vantaggio selettivo.

II.3 La dinamica Lotka-Volterra nei sistemi immunitari: preda-predatore tra CTL e virus

La dinamica dell'interazione tra linfociti T citotossici (CTL, il «predatore») e cellule infettate dal virus (la «preda») nei sistemi immunitari è formalizzata attraverso un sistema di equazioni di tipo Lotka-Volterra modificato. Sia $T(t)$ la concentrazione di cellule T infettate (target per i CTL), $V(t)$ la concentrazione di particelle virali libere, $E(t)$ la concentrazione di CTL virus-specifici, e $I(t)$ la concentrazione di cellule infettate (che producono nuovi virioni):

Il modello standard di infezione virale (Nowak e May, 2000) è:

$$dT/dt = \lambda - d_T \cdot T - \beta \cdot V \cdot T$$

$$dI/dt = \beta \cdot V \cdot T - \delta \cdot I - k \cdot E \cdot I$$

$$dV/dt = p \cdot I - c \cdot V$$

$$dE/dt = q \cdot E \cdot I - d_E \cdot E$$

dove λ è il tasso di produzione di cellule target sane, d_T è il tasso di morte delle cellule sane, β è la costante di infezione (il prodotto dell'incontro tra virus libero e cellule sane), δ è il tasso di morte delle cellule infettate, k è la costante di lisi citotossica (il tasso con cui i CTL uccidono le cellule infettate), p è la produzione virale per cellula infettata, c è il tasso di clearance virale, q è il tasso di espansione dei CTL in risposta all'antigene virale, e d_E è il tasso di morte dei CTL (o di declino dell'espansione clonale).

La dinamica all'equilibrio e le transizioni di fase. Il sistema ha due tipi di equilibri stabili, a seconda dei parametri:

Clearance virale: se il numero di riproduzione di base del virus $R_0 = \lambda\beta p / (d_T \cdot \delta \cdot c) < 1$, il virus viene eliminato e l'equilibrio è $(T^*, 0, 0)$ — nessuna infezione persistente.

Infezione cronica senza CTL efficaci: se $R_0 > 1$ ma la risposta CTL è insufficiente (k piccolo), l'equilibrio è una viremia stabile con alta carica virale.

Infezione cronica controllata dai CTL: se $R_0 > 1$ e la risposta CTL è efficace, l'equilibrio è una viremia bassa con alta risposta CTL — il pattern osservato nei long-term non-progressors dell'HIV.

Le oscillazioni nella fase acuta. Nella fase acuta dell'infezione virale, il sistema Lotka-Volterra immunitario non converge monotonicamente all'equilibrio ma attraversa oscillazioni smorzate. Il picco virale della fase acuta (quando la risposta CTL non è ancora pienamente sviluppata) è seguito da una caduta rapida della viremia (la risposta CTL è ormai attivata e abbatte le cellule infettate) e poi da un rimbalzo (il virus evolve varianti di escape CTL, o la risposta CTL si riduce per esaurimento), producendo la caratteristica curva «bimodale» della viremia nella fase acuta osservata in HIV, SARS-CoV-2, influenza e in molte altre infezioni virali.

Il periodo delle oscillazioni è determinato dagli stessi parametri del sistema Lotka-Volterra classico, con il prodotto $k \cdot q$ come il parametro di accoppiamento preda-predatore. Linearizzando il sistema attorno all'equilibrio di infezione cronica (I^*, E^*), le frequenze proprie delle oscillazioni sono:

$$\omega \approx \sqrt{(k \cdot q \cdot I \cdot E - (\delta + k \cdot E^*)^2/4)^{**}}$$

Per parametri tipici dell'infezione da HIV, $\omega \approx 0.1-0.5$ /settimana, producendo oscillazioni con periodicità di 2-10 settimane — coerentemente con le fluttuazioni osservate nella viremia in pazienti con HIV non trattato.

L'HIV come caso studio di coevoluzione immuno-virale. L'HIV è il caso più documentato di coevoluzione tra un virus e il sistema immunitario dell'ospite. Il virus evolve continuamente per sfuggire alla risposta CTL virus-specifica attraverso mutazioni nei peptidi presentati dall'MHC-I. La tasso di sostituzione nucleotidica dell'HIV (circa 3×10^{-5} sostituzioni per base per ciclo di replicazione) è di diversi ordini di grandezza superiore a quello del genoma dell'ospite — il che garantisce che il virus possa esplorare l'intero spazio delle varianti di escape nei peptidi presentati dall'MHC in pochi mesi di infezione.

Il modello della co-evoluzione HIV-CTL è un sistema di equazioni accoppiate per la dinamica dei quasispecies virali. Sia $v_j(t)$ la concentrazione del ceppo virale j e ϵ_j la sua sensibilità alla risposta CTL dell'ospite (basso ϵ_j significa alta resistenza — escape). Il sistema è:

$$dv_j/dt = r_j \cdot v_j \cdot (1 - \epsilon_j \cdot E) + \sum_k \mu\{kj\} \cdot v_k - d \cdot v_j$$

$$dE/dt = \alpha \cdot (\sum_j \epsilon_j \cdot v_j) \cdot E - \beta_E \cdot E$$

dove r_j è il tasso di replicazione del ceppo j , $\mu_{\{kj\}}$ è il tasso di mutazione del ceppo k verso il ceppo j (la struttura dello spazio delle mutazioni), e α è il tasso di espansione CTL proporzionale alla viremia pesata per la sensibilità. La coevoluzione produce un equilibrio dinamico in cui i ceppi virali di escape (basso ϵ_j) si accumulano ma hanno fitness di replicazione ridotta (r_j basso per i mutanti di escape, che spesso hanno difetti nelle funzioni proteiche), producendo il caratteristico set-point virale dell'HIV — una viremia stabile a circa 10^4 – 10^5 copie/ml che riflette l'equilibrio tra la pressione selettiva dei CTL e i costi fitness delle mutazioni di escape.

III. La sorveglianza immunologica e i limiti della verifica

La corsa agli armamenti immunitaria pone un problema strategico di eccezionale delicatezza: il sistema immunitario deve riconoscere ed eliminare i patogeni (non-self) senza attaccare le cellule proprie dell'organismo (self) — il problema della *tolleranza immunologica* già analizzato nella Monografia XXIII. La struttura del missing self recognition delle cellule NK — il sistema di verifica a default negativo $A(c) - I(c) > \theta_{NK}$ — e il processo dell'immunoediting (eliminazione → equilibrio → evasione) completano il quadro della sorveglianza come sistema di teoria dei giochi a tre fasi.

IV. Il cervello sociale e l'ipotesi di Dunbar: la coevoluzione cognitivo-sociale

IV.1 La correlazione neocorteccia-gruppo

Robin Dunbar (1992, 1998, 2009) ha proposto l'*ipotesi del cervello sociale* (*social brain hypothesis*, SBH) a partire da una correlazione empirica robusta: nel clade dei primati, la dimensione relativa della neocorteccia è correlata positivamente con la dimensione media del gruppo sociale della specie. La regressione di Dunbar (1992) su 38 specie di primati dà:

$$\log(N_{\text{sociale}}) = a + b \cdot \log(NQ)$$

con il quoziente neocorticale $NQ = \text{volume neocorticale} / \text{volume del resto del cervello}$. Il coefficiente $b \approx 3.8$ e $R^2 \approx 0.76$. Applicando questa regressione agli esseri umani ($NQ \approx 4.1$): $N_{\text{sociale}} \approx 148$ — il numero di Dunbar.

IV.2 La struttura gerarchica della rete sociale: i cerchi concentrici e la loro base neurale

Le reti sociali umane mostrano una struttura gerarchica di cerchi concentrici con scala approssimativa di 3:

$$n_k \approx 3^k \cdot n_0$$

con $n_0 \approx 1.5$ (il cerchio di massima intimità), $n_1 \approx 5$ (gruppo di supporto), $n_2 \approx 15$ (gruppo simpatico), $n_3 \approx 50$ (gruppo di affinità), $n_4 \approx 150$ (numero di Dunbar), $n_5 \approx 500$ (conoscenti), $n_6 \approx 1500$ (persone riconoscibili per nome).

La base neurale di questa struttura gerarchica è stata parzialmente identificata da Lewis et al. (2011) con una meta-analisi di studi neuroimaging sulla mentalizzazione e la elaborazione sociale: la corteccia prefrontale mediale (mPFC) è attivata in relazione alla *distanza sociale* — più vicino il legame sociale, più alta l'attivazione mPFC — e la giunzione temporoparietale (TPJ) è attivata principalmente per i legami nel cerchio di n_2 (15 persone), suggerendo che queste strutture implementino la mentalizzazione per le relazioni socialmente più rilevanti.

IV.3 Il costo cognitivo della mentalizzazione di ordine elevato

La capacità di mentalizzazione di ordine k ha costo computazionale che cresce esponenzialmente con k :

$$C(k) \approx b^k \cdot C(1)$$

dove $b > 1$ è il fattore di espansione per livello. Per $b \approx 1.5$ e $C(1) \approx 1$ slot di memoria di lavoro, $C(5) \approx 7.6$ — al limite della capacità della memoria di lavoro umana (circa 4 ± 1 slot secondo Cowan, 2001). Questo stima suggerisce che il limite a 5–6 ordini intenzionali osservato negli esseri umani riflette un vincolo computazionale reale, non una carenza evolutiva.

V. Il costo cognitivo del grooming e la derivazione del linguaggio

La derivazione quantitativa di Dunbar: sia τ_g il tempo dedicato al grooming fisico per mantenere un legame con un individuo (circa 30 minuti al giorno per i legami di massima intimità). La dimensione del gruppo gestibile con grooming fisico in un budget temporale di $\tau_{\max}$ è:

$$N_{\text{grooming}} \approx \tau_{\text{max}} / \tau_{\text{g}} \approx 6 \text{ ore} \cdot (1/0.5 \text{ ore}) = 12 \text{ individui}$$

Per raggiungere $N \approx 150$, il grooming verbale — il linguaggio — deve essere almeno 12 volte più efficiente per unità di tempo, il che è plausibile se il linguaggio permette di mantenere legami con 3–5 individui contemporaneamente durante una conversazione di gruppo.

Dunbar (1996) ha stimato che circa il 60–70% delle conversazioni quotidiane degli esseri umani riguarda informazioni sulle relazioni sociali — informazioni su chi fa cosa con chi, sulle reputazioni e le intenzioni degli assenti. Questa prevalenza del gossip nelle conversazioni umane riflette la funzione primaria del linguaggio come strumento di gestione della rete sociale piuttosto che come trasmissione di informazioni tecniche o ecologiche — una predizione confermata da studi di conversazione in contesti naturalistici in culture diverse (Dunbar, Marriott e Duncan, 1997).

VI. La circuiteria della dopamina sociale e i giochi di reputazione

La componente neurobiologica del gioco delle coalizioni sociali è implementata in larga misura attraverso il sistema dopaminergico mesolimbico. Le ricompense sociali (cooperazione, riconoscimento, approvazione sociale) attivano il nucleo accumbens e la corteccia orbitofrontale con la stessa struttura dell'errore di predizione del rinforzo identificata da Schultz per le ricompense materiali. Il valore sociale è, nel senso neurale rilevante, un valore — non una categoria distinta dall'utilità monetaria ma un'istanza della stessa architettura di valutazione.

La reputazione — il valore atteso delle interazioni future con un individuo, condizionato alle informazioni disponibili sul suo comportamento passato — è un oggetto che il cervello deve stimare in modo continuo per tutti i membri della propria rete sociale. La stima della reputazione richiede la memoria delle interazioni passate (il ruolo della memoria ippocampale episodica e semantica), la proiezione delle intenzioni future (la mentalizzazione di ordine superiore) e la valutazione affettiva della relazione (il ruolo dell'amigdala e dell'insula nella modulazione del valore sociale). Queste componenti non sono separate ma integrate in un sistema di valorizzazione sociale distribuito.

VII. La teoria dei giochi coalizionali e il valore di Shapley

La gestione delle coalizioni sociali — chi si allea con chi, contro chi, per ottenere quali risorse — è il problema strategico centrale. Un gioco coalizionale è definito da un insieme di giocatori $N = \{1, 2, \dots, n\}$ e da una funzione caratteristica $v : 2^N \rightarrow \mathbb{R}$. Il *valore di Shapley* — la soluzione razionale per distribuire $v(N)$ — è:

$$\varphi_i(v) = \sum_{\{S \subseteq N \mid i \in S\}} \left[\frac{|S|! \cdot (n - |S|)!}{n!} \right] \cdot [v(S \cup \{i\}) - v(S)]$$

che calcola il contributo marginale medio del giocatore i su tutte le possibili sequenze di formazione della grande coalizione. Il calcolo richiede la valutazione di 2^n sottoinsiemi — per $n = 150$, $2^{150} \approx 10^{45}$. La gestione strategica di coalizioni reali non può calcolare i valori di Shapley esatti ma deve utilizzare approssimazioni euristiche — e qui risiede la pressione selettiva che ha guidato l'espansione della neocorteccia.

VIII. Il limite della strategia evolutivamente stabile: perché il gioco non converge

In molti giochi coevolutivi biologicamente rilevanti — la corsa agli armamenti immunitaria, il gioco delle coalizioni sociali, la competizione per il rango — le *strategie evolutivamente stabili* (ESS) non esistono o non sono computazionalmente individuabili in tempo polinomiale. La non-esistenza dell'ESS produce dinamiche cicliche — la stessa struttura dell'impossibilità di convergenza al Nash nei giochi somma-zero discussa nella Monografia XIII — che si manifestano biologicamente come le oscillazioni della Red Queen, i cicli di dominanza nelle gerarchie di rango, e le fluttuazioni nella composizione delle coalizioni sociali.

Nella corsa agli armamenti HLA-patogeni, questa non-convergenza si manifesta nel polimorfismo persistente degli aplotipi HLA: nessun aplotipo può fissarsi perché non appena diventa comune i patogeni co-evolvono per sfuggirgli. Nella dinamica CTL-HIV, si manifesta nelle oscillazioni della viremia e nella diversificazione dei quasispecies virali. In entrambi i casi, il gioco non converge a un equilibrio stabile ma rimane in moto perpetuo — una dinamica di Red Queen che è il prodotto necessario della struttura del gioco, non un difetto del sistema evolutivo.

Questa non-convergenza non è un fallimento del sistema biologico: è la struttura formale dell'adattamento in ambienti coevolutivi. Un sistema biologico che convergesse a una strategia stabile sarebbe infinitamente vulnerabile all'innovazione strategica del proprio avversario. La mancata convergenza — la continua esplorazione e ri-ottimizzazione del paesaggio strategico — è il prezzo evolutivamente ottimale per mantenere la capacità di risposta alle innovazioni altrui. Il cervello sociale non è il prodotto di un gioco che ha trovato il suo equilibrio: è il prodotto di un gioco che non si ferma mai.

Monografia XXIII

La verifica biologica

Come il sistema biologico verifica i propri processi: il sistema immunitario come verificatore dell'identità cellulare, l'apoptosi come rimozione formale di computazioni errate, la DNA repair machinery, l'Unfolded Protein Response, l'autofagia e la biologia dei telomeri come sistemi di error correction biologico

I. Il paradosso della verifica interna: Gödel nei sistemi viventi

Nel 1931, Kurt Gödel dimostrò che nessun sistema formale sufficientemente potente può dimostrare la propria consistenza dall'interno. Qualsiasi tentativo di verificare dall'interno che il sistema non produca contraddizioni conduce o a incompletezza o a inconsistenza. Il teorema di Rice ne è la versione computazionale: non esiste un algoritmo generale per verificare proprietà semantiche non banali di programmi arbitrari.

I sistemi biologici affrontano questo stesso problema strutturale: le cellule devono verificare la correttezza dei propri processi — il DNA non è danneggiato, le proteine si ripiegano correttamente, le divisioni cellulari avvengono secondo il programma appropriato — ma il verificatore è esso stesso un sistema biologico soggetto agli stessi tipi di errore che deve rilevare. La biologia ha risposto a questa impossibilità gödelliana con sistemi di sorveglianza multipli, ridondanti e parzialmente indipendenti, ciascuno dei quali verifica aspetti complementari della correttezza cellulare senza che nessuno di essi

verifichi se stesso. Questa monografia analizza cinque sistemi di verifica biologica distinti e complementari.

II. La DNA repair machinery: error correction nel codice biologico

II.1 La struttura degli errori nel DNA e la loro frequenza

Il genoma dei mammiferi — circa 3×10^9 paia di basi — è soggetto a un flusso continuo di danni chimici. Le stime di Lindahl (1993, Premio Nobel 2015) indicano che ogni cellula subisce circa 10.000–20.000 lesioni al DNA al giorno da quattro fonti principali: l'idrolisi spontanea (depurinazione: perdita di guanina o adenina a circa 5.000 eventi/giorno/cellula; deaminazione: conversione della citosina in uracile a circa 100–500 eventi/giorno), le specie reattive dell'ossigeno (circa 10.000 ossidazioni di guanina in 8-oxoguanina/giorno), gli agenti alchilanti endogeni (S-adenosilmetionina che alchila spontaneamente le basi) e le rotture del doppio filamento (circa 10/giorno/cellula).

La frequenza di errore intrinseca della DNA polimerasi è dell'ordine di 10^{-5} per base per replicazione. Dopo la correzione esonucleasica (proofreading) incorporata nella polimerasi, la frequenza scende a circa 10^{-7} . Il sistema di mismatch repair (MMR) aggiuntivo porta la frequenza finale a circa 10^{-10} per base per replicazione:

$$f_{\text{finale}} = f_{\text{pol}} \times r_{\text{proof}} \times r_{\text{MMR}} \approx 10^{-5} \times 10^{-2} \times 10^{-3} = 10^{-10}$$

I meccanismi molecolari includono il *Nucleotide Excision Repair* (NER) per le lesioni che distorcono la doppia elica (dimeri di pirimidina UV-indotti, addotti degli idrocarburi policiclici aromatici), il *Base Excision Repair* (BER) per basi danneggiate di piccole dimensioni (uracile, 8-oxoguanina, basi alchilate) attraverso glicosilasi specifiche che riconoscono e flippano le basi danneggiate fuori dalla doppia elica, e il *Mismatch Repair* (MMR) che corregge gli errori di appaiamento prodotti dalla DNA polimerasi. La risposta al danno del DNA culmina nel sistema di checkpoint centrato su p53 — con la codifica dinamica oscillazione→arresto vs rampa→apoptosi descritta nella versione precedente di questa monografia.

III. La risposta alle proteine non ripiegate (UPR): error correction del proteoma

III.1 Il problema del controllo qualità delle proteine

Ogni cellula sintetizza circa 10.000 proteine diverse, ciascuna delle quali deve assumere la propria struttura tridimensionale nativa per essere funzionale. Il processo di ripiegamento è termodinamicamente guidato verso la struttura a energia libera minima (il teorema di Anfinsen, 1973: la struttura nativa è determinata dalla sequenza aminoacidica), ma è cinematicamente complesso: esistono numerose strutture intermedie metastabili che possono accumularsi come proteine mal ripiegate (*misfolded proteins*). Le proteine mal ripiegate sono pericolose perché tendono ad aggregarsi — le superfici idrofobiche normalmente sepolte nel core proteico vengono esposte all'esterno, producendo interazioni promiscue con altre proteine che generano aggregati amorfi o fibrille amiloidi (la struttura patologica delle proteinopatie neurodegenerative: malattia di Alzheimer, malattia di Parkinson, SLA).

Il *reticolo endoplasmatico* (RE) è la sede della sintesi e del ripiegamento delle proteine secretorie e di membrana — circa il 30% del proteoma cellulare. Il RE ha un sistema di controllo qualità strutturato: le proteine malripiegate vengono riconosciute da chaperone molecolari (BiP/GRP78, calreticulina, calnexina) che le tengono in soluzione prevenendo l'aggregazione, tentano di assisterle nel ripiegamento corretto, e se il ripiegamento fallisce le indirizzano alla degradazione (ERAD — ER-associated degradation — attraverso il sistema ubiquitina-proteasoma).

Quando la capacità del sistema di controllo qualità del RE è saturata — quando il numero di proteine malripiegate supera la capacità dei chaperone disponibili — si attiva la *Unfolded Protein Response* (UPR): una risposta trascrizionale e trasduzionale che segnala lo stato di «stress del RE» al nucleo e coordina un programma adattativo per ripristinare l'omeostasi proteica.

III.2 I tre sensori dell'UPR: IRE1, PERK e ATF6

L'UPR è attivata da tre sensori transmembrana del RE — IRE1 α (Inositol-Requiring Enzyme 1 α), PERK (PKR-like ER Kinase) e ATF6 (Activating Transcription Factor 6) — ciascuno con meccanismo di attivazione e effetti a valle distinti:

IRE1 α è una proteina transmembrana con dominio luminale rivolto verso il lume del RE (che rileva le proteine malripiegate attraverso interazione diretta o competizione con BiP) e due domini citosolici: un dominio kinasico e un dominio endoribonucleasico (RNasi). Nelle condizioni normali, IRE1 α è tenuto inattivo dall'associazione con BiP. Quando le proteine malripiegate si accumulano nel lume del RE, competono con IRE1 α per il legame a BiP; BiP si dissocia da IRE1 α ,

che oligomerizza e si trans-autofosforila (attivazione del dominio chinasi). L'attivazione della RNasi di IRE1 α produce due effetti:

Il primo è lo *splicing non convenzionale* dell'mRNA di XBP1 (X-Box Binding Protein 1): IRE1 α taglia l'mRNA di XBP1 in due siti specifici, rimuovendo un introne di 26 nucleotidi. Il frame-shift produce XBP1s (XBP1 spliced) — un fattore trascrizionale attivo che attiva centinaia di geni UPR target: i geni dei chaperone del RE (BiP, calreticulina, PDI), i geni della biogenesi del RE, e i geni dell'ERAD. Formalmente, XBP1s è il master regulator della risposta adattativa dell'UPR — l'analogo del master regulator MyoD per la differenziazione muscolare (Monografia XVI): la sua attivazione «accende» un intero programma trascrizionale coordinato.

Il secondo effetto della RNasi di IRE1 α è la *RIDD* (Regulated IRE1-Dependent Decay): la degradazione selettiva di centinaia di mRNA che codificano proteine del RE — una riduzione globale del carico di nuove proteine che devono essere ripiegate nel RE, alleggerendo la pressione sullo stress.

PERK è una chinasi transmembrana del RE che, quando attivata (per lo stesso meccanismo di competizione BiP/proteine malripiegate), fosforila il fattore di inizio traduzione eIF2 α (eukaryotic Initiation Factor 2 α) sulla serina 51. La fosforilazione di eIF2 α inibisce la traduzione globale — riduce il carico di nuove proteine sul RE di circa il 30–50% nella fase acuta dello stress — producendo il tempo necessario perché i chaperone possano processare le proteine malripiegate accumulate. Contemporaneamente, l'inibizione globale della traduzione paradossalmente aumenta la traduzione di specifici mRNA con uORF (upstream Open Reading Frames) nel loro 5'UTR — tra cui ATF4, un fattore trascrizionale che attiva i geni della risposta integrata allo stress (ISR): i geni dei chaperone, i geni dell'aminoacido synthesis (per produrre le materie prime proteiche), e — se lo stress è prolungato e severo — il gene del fattore apoptotico CHOP (C/EBP Homologous Protein).

ATF6 è un fattore trascrizionale legato alla membrana del RE: nelle condizioni normali è trattenuto nel RE per interazione con BiP; in condizioni di stress del RE, viene rilasciato dal BiP e trasportato all'apparato di Golgi, dove viene tagliato da due proteasi (S1P e S2P) liberando il suo dominio trascrizionale citosolico (ATF6f). ATF6f migra al nucleo e attiva i geni dei chaperone del RE e i geni della biogenesi del RE — una risposta trascrizionale che aumenta la capacità di ripiegamento del RE adeguandola al carico aumentato.

III.3 La decisione UPR: adattamento versus apoptosi

Il fattore trascrizionale CHOP — indotto dalla via PERK-eIF2 α -ATF4 in condizioni di stress del RE prolungato o severo — è il punto di incrocio tra la risposta adattativa e la risposta apoptotica dell'UPR. CHOP attiva:

I geni dell'apoptosi: DR5 (Death Receptor 5), BIM (un membro pro-apoptotico della famiglia Bcl-2 che inibisce la sopravvivenza). CHOP reprime contemporaneamente i geni anti-apoptotici come Bcl-2.

Il programma di stress ossidativo: CHOP aumenta la produzione di ROS nel RE attraverso l'attivazione di ERO1 (ER Oxidase 1) — le ROS eccessive danneggiano le proteine del RE e le membrane, accelerando l'ingresso in apoptosi.

La struttura formale della decisione UPR tra adattamento e apoptosi è analoga alla decisione p53 nella risposta al danno del DNA: il pattern temporale del segnale — la *cinetica* dell'attivazione dei tre sensori (IRE1, PERK, ATF6) — determina l'esito più della sua ampiezza. Lo stress moderato e transitorio produce attivazione rapida di IRE1 (XBP1s → risposta adattativa) e PERK (riduzione temporanea della traduzione) con recupero; lo stress severo e prolungato produce attivazione persistente di PERK con accumulo progressivo di CHOP e ingresso in apoptosi. Bernales, McDonald e Walter (2006) hanno documentato questa dipendenza dalla cinetica nelle cellule in coltura: la stessa dose di agente stressante produce adattamento se somministrata in modo cronico (le cellule si adattano gradualmente) e apoptosi se somministrata acutamente (le cellule non hanno tempo di adattarsi).

L'UPR è rilevante per numerose patologie: le malattie neurodegenerative (Alzheimer, Parkinson, SLA) sono associate a stress del RE cronico per l'accumulo di aggregati proteici che saturano il sistema ERAD; il diabete di tipo 2 è associato a stress del RE nelle cellule β pancreatiche per l'overload della sintesi di insulina in risposta all'iperglicemia; e le malattie infiammatorie croniche intestinali (IBD) sono associate a disfunzione dell'UPR nelle cellule caliciformi intestinali che producono grandi quantità di mucine glicosilate.

IV. L'autofagia: degradazione delle componenti cellulari danneggiate

IV.1 La struttura dell'autofagia

L'*autofagia* (dal greco *αὐτο-φαγεῖν*, mangiare se stessi) è il sistema cellulare di degradazione selettiva e non-selettiva di componenti cellulari danneggiate o superflue: organelli disfunzionali, aggregati proteici, patogeni intracellulari. L'autofagia produce l'encapsulamento del materiale da degradare in una doppia membrana (l'*autofagosoma*) che poi si fonde con i lisosomi, dove il contenuto viene degradato dalle idrolasi acide lisosomiali.

Esistono tre forme principali di autofagia:

Macroautofagia (semplicemente «autofagia» nella letteratura comune): formazione de novo di un autofagosoma attorno al materiale da degradare attraverso un processo di nucleazione, elongazione e chiusura della membrana. È la forma quantitativamente più rilevante e la meglio caratterizzata molecolarmente.

Mitofagia: una forma selettiva di macroautofagia che degrada selettivamente mitocondri disfunzionali — quelli con potenziale di membrana ridotto, con DNA mitocondriale mutato o con produzione eccessiva di ROS. La mitofagia è il meccanismo cellulare primario per il controllo qualità del compartimento mitocondriale e per la prevenzione dell'accumulo di mitocondri disfunzionali nel corso dell'invecchiamento.

Aggreffagia: una forma selettiva di macroautofagia che degrada selettivamente gli aggregati proteici — particolarmente gli aggregati ubiquitinati che si accumulano quando il sistema ubiquitina-proteasoma è saturo. L'aggreffagia è il meccanismo di «ultima risorsa» per la clearance degli aggregati proteici nelle malattie neurodegenerative.

IV.2 Il meccanismo molecolare: la chinasi ULK1 e il complesso PI3K-Beclin1

L'iniziazione dell'autofagia è controllata da due complessi proteici ad attività chinastica che fungono da interruttori molecolari:

Il *complesso ULK1* (Unc-51-Like Autophagy Activating Kinase 1) — che comprende ULK1, ATG13, FIP200 e ATG101 — è il punto di integrazione dei segnali nutrizione-status della cellula. Nelle condizioni di abbondanza di nutrienti, ULK1 è inibita dalla chinasi mTORC1 (mechanistic Target Of Rapamycin Complex 1), che fosforila ULK1 (S758) e ATG13 (S258) mantenendo il complesso inattivo. Nelle condizioni di stress nutrizionale, energetico o di stress del RE, mTORC1 si dissocia dal complesso ULK1 e la chinasi AMPK (AMP-activated Protein Kinase) — il sensore del deficit energetico — fosforila ULK1 (S317 e S777) attivandola. ULK1 attivata fosforila i

propri substrati nel complesso — ATG13 e FIP200 — potenziando la propria attività autofosforilante (feedback positivo), e fosforila i componenti del complesso Beclin1-PI3K che inizia la nucleazione della membrana del fagoforo.

Il *complesso PI3K-Beclin1* (che comprende VPS34, Beclin1, ATG14, p150 e AMBRA1) catalizza la produzione di fosfatidilinositolo-3-fosfato (PI3P) nelle membrane del reticolo endoplasmatico o delle membrane associate al RE (MAM — Mitochondria-Associated Membranes). Il PI3P recluta effettori FYVE-domain e PX-domain (WIPI1/2, DFCP1) che marcano il sito di nucleazione del fagoforo e guidano il reclutamento delle proteine della macchina di elongazione della membrana.

L'elongazione della membrana del fagoforo è mediata da due sistemi di coniugazione analoghi all'ubiquitinazione:

Il sistema ATG12-ATG5-ATG16L1: ATG12 viene coniugata covalentemente ad ATG5 (attraverso la reazione E1-like dell'ATG7 e la reazione E2-like dell'ATG10), e il complesso ATG12-ATG5 si associa con ATG16L1 per formare un complesso quaternario che localizza sulla membrana esterna del fagoforo e catalizza la coniugazione del secondo sistema.

Il sistema LC3-PE: LC3 (Microtubule-Associated Protein 1 Light Chain 3) viene processata dalla cisteina proteasi ATG4 (producendo LC3-I), poi attivata dall'E1-like ATG7, trasferita all'E2-like ATG3, e coniugata alla fosfatidiletanolamina (PE) nella membrana del fagoforo (producendo LC3-II). LC3-II è il marcatore standard della membrana autofagosomale — l'anticorpo anti-LC3 in immunofluorescenza produce il caratteristico pattern puntato («puncta») che visualizza gli autofagosomi nelle cellule in autofagia attiva.

La *chiusura* del fagoforo e la formazione dell'autofagosoma completo è mediata dall'ESCRT (Endosomal Sorting Complexes Required for Transport) e dall'attività ATG4 (che de-coniuga LC3-II dalla membrana esterna dell'autofagosoma, mantenendo LC3-II soltanto nella membrana interna — quella che rimarrà dopo la fusione con i lisosomi e la digestione del contenuto).

IV.3 L'autofagia selettiva: i recettori cargo e la specificità della degradazione

La macroautofagia non selettiva sequestra materiale citoplasmatico in modo promiscuo. L'autofagia selettiva (mitofagia, aggreffagia, xenofagia)

degrada specifici substrati attraverso *recettori cargo* — proteine adattatrici che riconoscono il substrato (mediante motivi di interazione con ubiquitina o con segnali specifici del substrato) e lo collegano alla membrana del fagoforo (mediante il loro LIR motif — LC3-interacting Region — che si lega direttamente a LC3-II). I principali recettori cargo nei mammiferi sono p62/SQSTM1, NBR1, NDP52, OPTN (optineurina) e TAX1BP1.

Per la mitofagia: il percorso PINK1-Parkin è il più studiato. La chinasi PINK1 (PTEN-Induced Kinase 1) viene normalmente importata nel mitocondrio sano e degradata dalla proteasi mitocondriale PARL. Nei mitocondri con potenziale di membrana ridotto (disfunzionali), PINK1 si accumula sulla membrana esterna mitocondriale (OMM) e si autofosforila (attivazione). PINK1 fosforila poi l'ubiquitina (S65) reclutando la ubiquitina ligasi E3 Parkin, che ubiquitina multipli substrati sulla OMM. Il segnale di ubiquitinazione viene riconosciuto dai recettori cargo (p62, OPTN, NDP52) che reclutano LC3-II e l'autofagosoma sul mitocondrio disfunzionale.

La rilevanza patologica della mitofagia è enorme: mutazioni loss-of-function in PINK1 e Parkin causano la malattia di Parkinson a insorgenza giovanile (PARK2 e PARK6), dimostrando che il controllo qualità mitocondriale mediato dalla mitofagia è essenziale per la sopravvivenza dei neuroni dopaminergici della substantia nigra. L'accumulo di mitocondri disfunzionali in assenza di mitofagia produce elevata produzione di ROS mitocondriali, attivazione dell'infiammazione (attraverso il rilascio di DNA mitocondriale nel citoplasma che attiva i recettori dell'immunità innata cGAS-STING) e disfunzione energetica — la combinazione che porta alla morte neuronale nella malattia di Parkinson.

V. La biologia dei telomeri: il limite strutturale della replicazione cellulare

V.1 Il problema della replica delle estremità e l'accorciamento telomerico

I telomeri sono strutture specializzate alle estremità dei cromosomi lineari dei eucarioti: sequenze ripetute di DNA non codificante (TTAGGG · n nelle cellule umane, con $n \approx 2.000\text{--}15.000$ ripetizioni a seconda dell'età e del tipo cellulare) associate a un complesso proteico protettivo (la *shelterina*, che comprende TRF1, TRF2, POT1, RAP1, TIN2 e TPP1) che protegge le estremità cromosomiche dal riconoscimento come DSB (rotture del doppio filamento) da parte dei

checkpoint del DNA, previene le fusioni end-to-end tra cromosomi, e regola la lunghezza telomerica.

Il problema della replica delle estremità (*end-replication problem*, Watson, 1972) è strutturale: la DNA polimerasi può sintetizzare DNA soltanto in direzione 5'→3' e ha bisogno di un primer di RNA per iniziare la sintesi. Nella replicazione del filamento ritardato (Okazaki fragments), il primer dell'ultimo frammento di Okazaki viene rimosso ma non può essere rimpiazzato da DNA sintetizzato *de novo* (non c'è DNA upstream a cui agganciare la sintesi). Il risultato è che ogni ciclo di replicazione produce una perdita di 50–200 bp di DNA telomerico dal 3' del filamento ritardato.

Nel corso della vita di un organismo, questo accorciamento progressivo — circa 50–200 bp per divisione cellulare in cellule somatiche umane — porta infine le cellule a una crisi replicativa: quando la lunghezza dei telomeri scende sotto un valore critico (circa 5 kb nel telomero più corto), la shelterina non riesce più a proteggere le estremità cromosomiche, che vengono riconosciute come DSB dai sensori ATM/ATR. Questo produce l'attivazione persistente del checkpoint del DNA con p53 e l'entrata in *senescenza replicativa* — l'arresto irreversibile del ciclo cellulare descritto da Hayflick (1961).

V.2 Il limite di Hayflick e la senescenza replicativa

Leonard Hayflick (1961, 1965) osservò che le cellule diploidi umane in coltura — i fibroblasti polmonari fetali (cellule WI-38) — non si dividono indefinitamente: dopo circa 50 ± 10 divisioni, le cellule cessano di proliferare entrando in uno stato morfologicamente distinto dalla morte cellulare (non muoiono, ma rimangono metabolicamente attive senza dividersi ulteriormente) che Hayflick denominò fase III, poi rinominata *senescenza replicativa*.

Il *limite di Hayflick* — circa 50 divisioni per i fibroblasti fetali umani, con valori variabili da 20 a 100 tra tipi cellulari diversi — è la manifestazione comportamentale dell'accorciamento telomerico progressivo. La dimostrazione diretta del nesso causale è venuta da Harley, Futcher e Greider (1990) che hanno mostrato la correlazione inversa tra lunghezza dei telomeri e numero di divisioni residue nelle cellule in coltura, e da Bodnar et al. (1998) che hanno dimostrato che l'introduzione ectopica della telomerasi nei fibroblasti umani estende la vita replicativa delle cellule oltre il limite di Hayflick senza produrre immortalizzazione cancerosa.

La struttura formale dell'accorciamento telomerico è quella di un *contatore molecolare*.

$$L_n = L_0 - \Delta l \cdot n$$

dove L_n è la lunghezza del telomero più corto dopo n divisioni, L_0 è la lunghezza iniziale (circa 15 kb nei neonati), Δl è l'accorciamento per divisione (50–200 bp, con la stima più realistica $\Delta l \approx 100$ bp), e n è il numero di divisioni. La cellula entra in senescenza quando $L_n < L_{cr}$ dove $L_{cr} \approx 5$ kb è la lunghezza critica minima per la protezione shelterina.

Il numero massimo di divisioni è:

$$n_{max} = (L_0 - L_{cr}) / \Delta l \approx (15.000 - 5.000) / 100 = 100 \text{ divisioni}$$

Il limite di Hayflick osservato di circa 50 divisioni è inferiore a questa stima perché: i fibroblasti fetali hanno già subito alcune divisioni in vivo prima di essere posti in coltura, la distribuzione degli accorciamenti è stocastica (alcuni telomeri si accorciano più velocemente di altri), e la senescenza viene innescata dal telomero *più corto*, non dalla lunghezza media.

V.3 La telomerasi: la soluzione biologica al problema dell'end-replication

La *telomerasi* (scoperta da Blackburn e Greider nel 1984, Premio Nobel 2009 insieme a Szostak) è una ribonucleoproteina con attività di trascrittasi inversa specifica per la sintesi delle sequenze telomeriche. Essa è composta da:

TERT (Telomerase Reverse Transcriptase): la subunità catalitica con attività enzimatica di trascrittasi inversa. *TERT* usa la subunità RNA come stampo per l'aggiunta di ripetizioni telomeriche alle estremità cromosomiche.

TERC (Telomerase RNA Component, anche TR): la subunità RNA che contiene la sequenza template 3'-AAUCCC-5' complementare alla sequenza telomerica TTAGGG. *TERC* si allinea al 3'-overhang del telomero e *TERT* sintetizza la ripetizione TTAGGG, poi la ribonucleoproteina si trasla di una ripetizione e sintetizza un'altra ripetizione — un meccanismo di sintesi processiva che aggiunge centinaia di ripetizioni per sessione.

La telomerasi è attiva nelle cellule della linea germinale (mantenendo la lunghezza telomerica alle lunghezze germinali di ≈ 15 kb), nelle cellule

staminali embrionali e in alcune cellule staminali adulte, e nelle cellule tumorali (dove la riattivazione della telomerasi — che avviene nell'85–90% dei tumori umani — è un prerequisito per l'immortalizzazione neoplastica). Nelle cellule somatiche differenziate, l'espressione di TERT è soppressa, producendo l'accorciamento progressivo che porta alla senescenza replicativa.

La rilevanza dell'accorciamento telomerico per l'invecchiamento biologico è supportata da studi epidemiologici che mostrano che la lunghezza media dei telomeri nei leucociti del sangue periferico (misurata con qPCR o con FISH) è inversamente correlata con l'età cronologica, con la mortalità complessiva, con l'incidenza delle malattie cardiovascolari, e con molti indicatori di fragilità nell'anziano. La lunghezza dei telomeri è stata proposta come «marcatore biologico dell'invecchiamento» — una misura dell'«età biologica» che può differire dalla età cronologica.

Il meccanismo attraverso cui la senescenza replicativa mediata dall'accorciamento telomerico contribuisce all'invecchiamento organico è il *Senescence-Associated Secretory Phenotype* (SASP): le cellule senescenti non muoiono ma diventano metabolicamente iperattive in un senso dannoso per il tessuto circostante, secernendo alti livelli di citochine pro-infiammatorie (IL-6, IL-8, TNF- α), metalloproteinasi di matrice (MMP-1, MMP-3, MMP-10), e fattori di crescita che alterano il microambiente tissutale, promuovono la disfunzione delle cellule circostanti, e — paradossalmente — possono favorire la crescita delle cellule precancerose nel tessuto adiacente.

VI. La ridondanza come strategia di verifica: il principio di Szilárd applicato alla biologia

Il tema unificante dei cinque sistemi di verifica analizzati — DNA repair, UPR, autofagia, apoptosi, immunosorveglianza — è la *ridondanza*: nessun singolo meccanismo è sufficiente, e la resistenza all'errore emerge dalla sovrapposizione di sistemi multipli, parzialmente indipendenti, che verificano aspetti complementari della correttezza biologica.

Sia $P_{\text{err}}(k)$ la probabilità che un errore sfugga a tutti i k sistemi di verifica ridondanti, e sia p la probabilità che un singolo sistema di verifica non rilevi l'errore. Se i sistemi sono indipendenti:

$$P_{\text{err}}(k) = p^k$$

Con $k = 5$ sistemi e $p = 0.1$: $P_{\text{err}}(5) = 10^{-5}$. La ridondanza moltiplicativa permette di raggiungere accuratissime arbitrariamente alte con componenti imperfette.

Il collegamento con la termodinamica è fornito dal principio di Szilárd (1929) e dalla generalizzazione di Landauer: qualsiasi operazione logica irreversibile — qualsiasi operazione che cancella informazione — ha un costo entropico minimo di $k_B \cdot \ln(2)$ per bit cancellato. L'eliminazione di una cellula apoptotica — l'eliminazione di un'unità computazionale corrotta — è un'operazione irreversibile che cancella l'intera informazione contenuta nel genoma (circa 3×10^9 bit) e nell'epigenoma. Il costo termodinamico minimo dell'apoptosi è:

$$\Delta S_{\text{min}} = k_B \cdot \ln(2) \cdot (3 \times 10^9) \approx 2 \times 10^{-12} \text{ J/K}$$

L'energia dissipata dall'apoptosi reale — misurata calorimetricamente — è dell'ordine di 10^{-10} J per cellula, circa cento volte il limite di Landauer. Questo fattore 100 è il «costo di implementazione» che riflette la non-idealità dei meccanismi molecolari reali rispetto ai limiti teorici.

VII. I limiti strutturali della verifica biologica: l'invecchiamento come accumulo irriducibile

Il limite finale della verifica biologica è temporale: i sistemi di error correction riducono ma non eliminano la frequenza degli errori, e nel corso del tempo gli errori irrilevanti si accumulano. L'invecchiamento biologico è, in larga parte, il manifestarsi di questo accumulo — un processo formalizzato dagli hallmark dell'invecchiamento (López-Otín et al., 2013, 2023): instabilità genomica (accumulo di mutazioni somatiche irriparate), perdita dei telomeri (discutibile nella sezione precedente), disfunzione mitocondriale (accumulo di mutazioni nel DNA mitocondriale non protetto da istoni), perdita della proteostasi (accumulo di proteine mal ripiegate che saturano i sistemi ERAD, UPR e autofagosomale), inflammaging (attivazione cronica di basso grado del sistema immunitario per il riconoscimento dei DAMP rilasciati dalle cellule danneggiate o senescenti).

La struttura formale dell'invecchiamento come accumulo di errori è quella di un processo stocastico in cui la frequenza di errori irrilevanti ϵ_{irr} cresce nel tempo, e le conseguenze di quegli errori si accumulano in modo non lineare:

$$d[\text{danno}]/dt = r_{\text{danno}} - r_{\text{repair}} - r_{\text{elimination}} \cdot \epsilon_{\text{detect}}$$

dove r_{danno} è il tasso di produzione di danni, r_{repair} è la capacità di riparazione, $r_{\text{elimination}}$ è la capacità di eliminazione (apoptosi e senescenza), e ϵ_{detect} è l'efficienza di rilevamento. Con l'invecchiamento, r_{repair} diminuisce (la macchina di riparazione stessa si degrada), ϵ_{detect} diminuisce (il sistema immunitario diventa meno efficiente nella sorveglianza), e la telomerasi somatica è assente (producendo l'accorciamento progressivo che porta alla senescenza replicativa), producendo un accumulo accelerato di danno — una dinamica di feedback positivo che è la firma formale dell'invecchiamento come processo biologico irriducibile.

Questa irriducibilità è l'enunciato biologico del teorema di Rice nella sua versione temporale: non esiste un sistema di verifica biologica che mantenga indefinitamente l'integrità di un sistema vivente, perché il sistema di verifica stesso è soggetto agli stessi processi di degrado che deve verificare. Il verificatore non può verificare se stesso.

Monografia XXIV

Il corpo come indeterminazione

La biologia ai propri limiti: l'irriducibilità del corpo vissuto, il chiasmo di Merleau-Ponty, i neuroni specchio come substrato dell'*embodied simulation*, l'intercorporeità e la distanza strutturale tra meccanismo ed esperienza

I. Il paradosso finale: il corpo che studia se stesso

Il ciclo che qui si conclude ha attraversato ventitré stazioni — dalla geometria dell'apprendimento alla genetica dello sviluppo, dalla neurobiologia del valore ai giochi evolutivi dell'immunità, dalla struttura formale della memoria alla fisica dell'emozione. In ciascuna stazione, il metodo era invariante: costruire il modello formale, derivarne le proprietà, confrontarle con l'evidenza sperimentale, tracciare i limiti del modello e indicare la direzione che eccede i suoi confini. Il metodo funziona. Funziona così bene che una tentazione si forma lungo il percorso: la tentazione di concludere che esso sia sufficiente — che il corpo biologico, una volta descritto con sufficiente precisione e profondità formale, sia anche *compreso*.

Questa monografia conclusiva argomenta contro quella tentazione. Non per sfiducia nel metodo — ma perché esiste una proprietà del corpo vivente che il metodo non può catturare, non per difetti tecnici che potrebbero essere corretti, ma per ragioni strutturali che derivano dalla natura stessa dell'oggetto. Quella proprietà è l'*essere-corpo* — il fatto che il corpo biologico non è soltanto un oggetto da descrivere ma il soggetto di ogni descrizione, inclusa quella che di lui fa la neurobiologia. Il corpo che studia se stesso è il paradosso che questa monografia prende sul serio.

II. La radice kantiana: il corpo come condizione di possibilità

La tradizione trascendentale kantiana ha individuato con precisione la struttura del paradosso. Kant ha mostrato nella *Critica della ragione pura* che lo spazio, il tempo e le categorie dell'intelletto non sono strutture del mondo in sé, ma le forme *a priori* attraverso cui il soggetto organizza l'esperienza. Il soggetto trascendentale non è un oggetto nel mondo — è la condizione di possibilità della conoscenza del mondo. Non si può conoscere il soggetto trascendentale nello stesso modo in cui si conoscono gli oggetti, perché il soggetto trascendentale è il lato dal quale si guarda, non quello che viene guardato.

Il corpo vissuto occupa per la biologia una posizione strutturalmente analoga. Ogni misurazione, ogni esperimento, ogni costruzione di modello avviene attraverso il corpo del ricercatore — attraverso la percezione, l'attenzione, la memoria, l'intenzione. Il corpo vissuto è la condizione di possibilità dell'intera impresa scientifica. Ma allora il corpo che la neurobiologia descrive — la corteccia visiva, il sistema limbico, la gerarchia del predictive coding — non è il corpo come condizione, bensì il corpo come oggetto. E la distanza tra i due non è colmabile moltiplicando i dati, perché i dati aggiuntivi sono anch'essi raccolti attraverso il corpo come condizione.

Questo non è un argomento contro la neurobiologia: è un argomento per la lucidità epistemologica — per la consapevolezza del tipo di conoscenza che la neurobiologia produce e del tipo che strutturalmente non può produrre.

III. Il chiasmo di Merleau-Ponty: la reversibilità del sentire

La fenomenologia del corpo vissuto ha trovato la sua elaborazione più precisa nell'opera tarda di Maurice Merleau-Ponty — particolarmente nel manoscritto incompiuto *Le Visible et l'Invisible*, pubblicato postumo nel 1968, e nelle *Notes de travail* che lo accompagnano. Il concetto centrale di quell'opera è il *chiasmo* (dal greco *χίασμος*, incrocio) — la struttura di reversibilità che definisce il corpo come luogo in cui il sentire e il sentito non sono due entità separate ma i due versanti di una stessa realtà che si incrociano senza mai coincidere.

L'esempio fondamentale di Merleau-Ponty è quello della mano che tocca e viene toccata. Quando la mano destra tocca la mano sinistra, si può esperire alternativamente — non simultaneamente — il toccare (la mano destra come soggetto del tocco) e l'essere toccata (la mano sinistra come oggetto del tocco). I due ruoli sono reversibili: la mano sinistra può diventare soggetto e la destra oggetto. Questa reversibilità — il chiasmo tra toccante e toccato — non è un caso anatomico ma struttura il rapporto del corpo con il mondo in generale.

Il corpo vedente e il visibile non sono due regioni separate del reale — il soggetto mentale di fronte all'oggetto fisico — ma appartengono alla stessa *Chair* (Carne), il termine con cui Merleau-Ponty designa il medium comune del sentire e del sentito. La carne non è materia nel senso cartesiano (*res extensa*) né spirito (*res cogitans*): è l'elemento in cui si formano sia il corpo come soggetto sia le cose come oggetti, il tessuto nel quale la distinzione soggetto-oggetto è già sempre operante ma non assoluta, sempre attraversabile ma mai cancellabile.

Il chiasmo è la struttura di questa attraversabilità: il punto in cui il vedente diventa visibile, il toccante diventa toccabile, il sentiente diventa sentito — senza che questa reversibilità sia mai una coincidenza. Il corpo non si vede mai come le cose del mondo: c'è sempre una differenza tra come mi appare la mia mano e come appare la mano di un altro. La reversibilità è strutturale ma incompleta — e questa incompletezza è la struttura stessa del corpo come soggetto: un soggetto che è anche nel mondo, che ha la stessa texture del mondo che percepisce, ma che non coincide mai interamente con nessun oggetto di quel mondo.

La rilevanza di questa struttura per la neurobiologia è precisa: il chiasmo di Merleau-Ponty indica che il confine tra il corpo come oggetto della neuroscienza (la corteccia somatosensoriale, il sistema propriocettivo, le mappe corporee nell'area S1) e il corpo come soggetto dell'esperienza (il corpo vissuto, lo schema corporeo, l'intenzionalità motoria) non è un confine tra due livelli di descrizione che potrebbero in linea di principio essere unificati — è un confine strutturale, il confine stesso tra il lato dal quale si guarda e quello che viene guardato. La neuroscienza può mappare con grande precisione i correlati neurali dello schema

corporeo; non può mappare il fatto che quello schema è vissuto da qualcuno.

IV. I neuroni specchio come substrato dell'*embodied simulation*

IV.1 La scoperta di Rizzolatti e il meccanismo di risonanza motoria

Nel 1992, Giacomo Rizzolatti e i suoi collaboratori a Parma registravano l'attività dei neuroni nell'area premotoria ventrale (area F5) del macaco durante l'esecuzione di azioni di presa. Per caso, osservarono che alcuni neuroni di F5 scaricavano non soltanto quando la scimmia eseguiva la presa, ma anche quando *osservava* un essere umano eseguire la medesima azione. Questi neuroni — ribattezzati *neuroni specchio* (mirror neurons) — erano selettivi per la stessa azione: un neurone che scaricava per la presa di precisione (indice-pollice) scaricava anche quando osservava una presa di precisione, ma non durante la presa di potenza.

La caratteristica formale dei neuroni specchio è la *corrispondenza funzionale* tra l'azione eseguita e l'azione osservata: il neurone codifica lo stesso atto motorio indipendentemente dal fatto che sia eseguito o osservato, e questa equivalenza funzionale avviene senza che i neuroni ricevano input diretti dagli organi di senso. La corrispondenza è mediata da connessioni che collegano la rappresentazione visiva dell'azione (nell'area STS — sulcus temporalis superior — che riconosce il movimento biologico) con la rappresentazione motoria della stessa azione nell'area F5.

IV.2 Il sistema specchio negli esseri umani: basi anatomiche e metodologiche

Negli esseri umani, i neuroni specchio non possono essere registrati direttamente (la registrazione di singolo neurone richiede elettrodi intracerebrali usati soltanto in contesti clinici). L'evidenza per un sistema specchio umano viene da metodiche indirette: la TMS (stimolazione magnetica transcranica) sulla corteccia motoria primaria produce potenziali evocati motori (MEP) nei muscoli della mano significativamente più ampi quando il soggetto *osserva* un'azione della mano rispetto alla condizione di controllo — indicando che l'osservazione dell'azione facilita l'eccitabilità del sistema motorio come se l'azione venisse eseguita. L'EEG mostra la soppressione del ritmo mu

(8–13 Hz), che è il marcatore di attivazione del sistema motorio, durante l'osservazione di azioni biologiche. La fMRI mostra attivazione della corteccia premotoria ventrale (BA 44/45 — la regione omologa dell'area F5 del macaco, corrispondente alla parte posteriore dell'area di Broca) e del lobulo parietale inferiore (BA 40) durante sia l'esecuzione sia l'osservazione di azioni, con overlap significativo tra i due pattern di attivazione.

Le aree che mostrano questo pattern di doppia attività — durante sia l'esecuzione sia l'osservazione delle azioni — costituiscono il *sistema specchio umano* (HMS — Human Mirror System). Oltre all'area premotoria ventrale e al lobulo parietale inferiore, l'HMS include l'area motoria supplementare, la corteccia somatosensoriale secondaria (S2), e — per la componente emotiva — l'insula anteriore che mostra attivazione durante sia la prova soggettiva di disgusto sia l'osservazione del disgusto sul volto altrui (Wicker et al., 2003).

IV.3 Embodied simulation: comprendere l'altro attraverso il proprio corpo

Vittorio Gallese e collaboratori (Gallese, 2001, 2005; Gallese e Goldman, 1998) hanno proposto il concetto di *embodied simulation* — simulazione incarnata — come il meccanismo attraverso cui il sistema specchio implementa la comprensione dell'altro. La tesi è che la comprensione delle azioni, delle emozioni e delle sensazioni altrui non avvenga attraverso un processo di inferenza esplicita (nel senso di: osservo il comportamento → inferisco lo stato mentale che lo produce), bensì attraverso la risonanza motoria e affettiva — l'attivazione automatica, subpersonale, delle stesse rappresentazioni neurali che verrebbero attivate se il soggetto stesse eseguendo quell'azione o vivendo quell'emozione.

La struttura formale della embodied simulation è quella di un'inferenza analogica mediata dal corpo: per comprendere l'azione *a* dell'individuo *O*, il sistema nervoso del soggetto *S* attiva la propria rappresentazione motoria di *a* — la stessa rappresentazione che *S* userebbe per eseguire *a* — e ne legge lo stato come informazione sullo stato di *O*. Il meccanismo è:

$$\text{rappresentazione_S(a_O)} \approx \text{rappresentazione_S(a_S)}$$

La comprensione è mediata dalla simulazione: *S* comprende ciò che *O* fa perché il proprio sistema motorio si trova, momentaneamente, in uno stato funzionalmente equivalente a quello di *O* durante l'esecuzione di *a*.

Questo meccanismo è strettamente connesso al chiasmo di Merleau-Ponty: la reversibilità tra il toccante e il toccato, tra il vedente e il visibile, trova nel sistema specchio la propria base neurale. Il mio corpo comprende il corpo altrui non descrivendo le sue proprietà dall'esterno, ma risuonando con lui dall'interno — producendo una conoscenza che non è proposizionale («il mio interlocutore è in uno stato di paura») ma pratica, presimbolica, incarnata.

La rilevanza filosofica è che la embodied simulation offre un substrato neurale per quella che Merleau-Ponty chiama la comprensione *intercorporea* — la comprensione dell'altro attraverso il proprio corpo, non attraverso il ragionamento esplicito. Il dualismo cartesiano — io come *res cogitans* che inferisce l'esistenza di altre *res cogitantes* a partire dall'analogia con il proprio stato mentale — è sostituito da una struttura di *risonanza corporea diretta*: il mio sistema motorio risponde al corpo dell'altro come se fosse il proprio, e questa risposta è il veicolo primario della comprensione intersoggettiva.

V. L'intercorporeità: la costituzione sociale dello schema corporeo

La nozione di *intercorporeità* — la costituzione del corpo vissuto attraverso l'incontro con i corpi altrui — è un prolungamento della fenomenologia di Merleau-Ponty verso la dimensione intersoggettiva. Lo schema corporeo — la rappresentazione preriflessiva del proprio corpo come sistema di posture, movimenti e capacità d'azione — non è costruito da ciascun individuo in isolamento a partire dalla propria esperienza sensoriale e propriocettiva: è costituito, in modo fondamentale, attraverso l'incontro con altri corpi.

I meccanismi neurobiologici dell'intercorporeità sono molteplici:

L'imitazione neonatale (Meltzoff e Moore, 1977) — la capacità dei neonati nelle prime ore di vita di imitare espressioni facciali dell'adulto (protrusione della lingua, apertura della bocca) — è stata interpretata come la manifestazione più precoce del sistema specchio: il neonato possiede già, alla nascita, un collegamento funzionale tra la rappresentazione visiva del volto altrui e la rappresentazione motoria dei propri movimenti facciali. Questo collegamento — il neonato non ha mai visto il proprio volto e non può sapere per esperienza diretta che cosa corrisponde motorialmente alla protrusione della lingua che vede nell'adulto — non può essere appreso dall'esperienza individuale e deve essere preformato filogeneticamente. L'intercorporeità è, in questo senso, evolutivamente primitiva: il corpo umano è, dalla nascita, già «accordato» ai corpi altrui.

L'apprendimento motorio per imitazione nella prima infanzia costituisce il secondo meccanismo: le competenze motorie — la prensione, il cammino, l'uso degli utensili, le produzioni vocali — vengono acquisite prevalentemente attraverso l'osservazione e l'imitazione di modelli adulti. Il sistema specchio è il substrato di questo apprendimento per imitazione: l'osservazione dell'azione attiva la rappresentazione motoria corrispondente, e questa rappresentazione fornisce il punto di partenza per la pratica motoria propria. Lo schema corporeo si costruisce osservando e risuonando con i corpi altrui — è un prodotto intersoggettivo, non individuale.

La biologia del riconoscimento del movimento biologico (area MT/V5 e STS posteriore) offre il substrato neurale della percezione dell'altro come agente. I display di *point-light* di Johansson (1973) — in cui il corpo umano in movimento è ridotto a pochi punti luminosi sulle giunture — sono istantaneamente riconosciuti come movimento biologico da osservatori umani, anche quando la stessa configurazione di punti presentata in modo statico o invertito non viene riconosciuta. Questo riconoscimento è preriflessivo — avviene in meno di 200 ms, prima che qualsiasi ragionamento esplicito possa intervenire — e attiva selettivamente il STS posteriore e il sistema specchio. Il corpo altrui viene riconosciuto come corpo — come agente con intenzioni e capacità — attraverso un meccanismo corporeo, non attraverso un'inferenza proposizionale.

La rilevanza clinica dell'intercorporeità è visibile nelle sindromi di disturbo dello schema corporeo e nella neuropsicologia dell'autismo. Nella sindrome da sconnessione interemisferica (split-brain), ciascun emisfero mantiene uno schema corporeo parzialmente autonomo. Nel disturbo dello spettro autistico, le anomalie nel sistema specchio — ridotta soppressione del ritmo μ durante l'osservazione di azioni altrui, ridotta attivazione dell'area di Broca e del lobulo parietale inferiore durante i compiti di imitazione — sono state proposte come correlato neurale della difficoltà di comprensione intersoggettiva caratteristica del disturbo (Oberman et al., 2005; Dapretto et al., 2006).

VI. La teoria della mente e i suoi limiti neurologici

La *teoria della mente* (ToM) — la capacità di attribuire stati mentali a se stessi e agli altri — è stata discussa nella Monografia XIX in relazione all'origine del linguaggio. Qui interessa la sua relazione con il corpo come veicolo della comprensione intersoggettiva.

La classica psicologia delle ToM ha proposto un modello *teorico* della comprensione dell'altro: comprendere l'altro significa costruire una *teoria*

dei suoi stati mentali, inferendo intenzioni, credenze e desideri a partire dall'osservazione del comportamento. Il sistema specchio propone invece un meccanismo *simulazionista*: la comprensione dell'altro è mediata dalla simulazione incarnata, non dall'inferenza teorica.

La distinzione non è puramente filosofica: i due meccanismi fanno predizioni diverse sull'architettura neurale e sui deficit clinici. Il modello teorico predice che la comprensione dell'altro deve coinvolgere le aree di ragionamento esplicito (corteccia prefrontale dorsale, giunzione temporoparietale — TPJ). Il modello simulazionista predice che deve coinvolgere il sistema motorio e le aree dell'elaborazione sensoriale (il sistema specchio). I dati neuroimaging indicano che entrambe le componenti sono attive, con le aree del sistema specchio che mediano la comprensione delle azioni e degli stati emotivi, e le aree prefrontali e la TPJ che mediano la comprensione delle credenze e dei desideri.

La convergenza più recente — la posizione di *simulazione critica* di Gallagher (2007) e di *embodied cognition* di Varela, Thompson e Rosch (1991) — propone che i due meccanismi non siano alternativi ma complementari: la simulazione incarnata fornisce la comprensione di base, preriflessiva, delle azioni e degli stati emotivi altrui, mentre l'inferenza teorica esplicita viene impiegata nei casi più complessi — mentalizzazione di ordine superiore, attribuzione di credenze false, ragionamento controfattuale sugli stati mentali.

VII. Il problema mente-corpo e la distanza strutturale: Quine, Davidson e il loro limite

La tradizione analitica anglosassone ha affrontato il problema mente-corpo attraverso le dottrine del *materialismo non riduzionista*. W.V.O. Quine — nella sua posizione di *naturalismo* — ha proposto che la psicologia debba essere continua con la biologia: non ci sono entità mentali ontologicamente distinte, ma soltanto stati fisici del sistema nervoso descritti a livelli di astrazione diversi. Donald Davidson, con il suo *monismo anomalo*, ha raffinato questa posizione: ogni evento mentale è identico a un evento fisico (monismo), ma le leggi psicologiche non riducono alle leggi fisiche (anomalismo) — perché le proprietà mentali sono realizzate da molteplici substrati fisici diversi (multiple realizability) e perché il dominio mentale è governato da principi di razionalità (il *principio di carità*) che non hanno analoghi nelle leggi fisiche.

Il monismo anomalo di Davidson offre una struttura che sembra promettente per la neurobiologia: i processi neurali e i processi mentali sono identici a livello degli eventi token (ogni singolo evento mentale è identico a un singolo evento fisico), ma non a livello delle proprietà (le

proprietà mentali non sono identiche alle proprietà fisiche — non c'è una legge-ponte che connette ogni tipo di stato mentale a ogni tipo di stato fisico). Questo implica che una neurobiologia completamente esaustiva — che descrivesse ogni neurone, ogni sinapsi, ogni pattern di attivazione — descriverebbe ogni evento fisico che è identico a un evento mentale, senza per questo *spiegare* le proprietà mentali di quegli eventi.

Il divario esplicativo di Levine — l'incapacità della descrizione fisica, anche la più completa, di spiegare *perché* un dato processo fisico è accompagnato da una data esperienza soggettiva — è la versione contemporanea del punto dove sia la fenomenologia che la filosofia analitica arresta il proprio passo. Merleau-Ponty e Davidson convergono nel riconoscere che il corpo vissuto eccede la sua descrizione in terza persona — per ragioni diverse ma con la stessa conclusione: la distanza tra la biologia del corpo e il corpo come soggetto di esperienza è strutturale, non contingente.

VIII. L'indeterminazione come proprietà costitutiva del vivente

La filosofia della fisica ha elaborato la nozione di *indeterminazione* come proprietà del livello quantistico della realtà. Ma l'indeterminazione rilevante per la biologia — quella che questa monografia ha cercato di articolare — non è l'indeterminazione di Heisenberg (sebbene essa contribuisca alla stocasticità dei processi molecolari biologici): è un'indeterminazione di livello diverso, che riguarda la relazione tra la descrizione in terza persona del sistema biologico e la sua auto-comprensione in prima persona.

Il vivente è, per struttura, un sistema che si autodescrive. Ogni organismo con sistema nervoso riceve informazioni dall'ambiente attraverso i propri recettori, le elabora attraverso la propria architettura neurale, e produce risposte comportamentali che modificano l'ambiente e quindi i propri input futuri. Questo circolo — sensazione → elaborazione → azione → sensazione modificata — è la struttura generale dell'autopoiesi discussa nella Monografia II. Ma nei sistemi con sufficiente complessità corticale, questo circolo include la *rappresentazione di se stessi*: il sistema non soltanto elabora l'ambiente ma si rappresenta come elaboratore, costruisce modelli dei propri processi, anticipa le proprie reazioni e le modera in funzione di quella anticipazione.

Questa auto-rappresentazione produce un'indeterminazione strutturale: il sistema che si auto-describe modifica se stesso attraverso la descrizione. Non è l'indeterminazione del principio di Heisenberg (la

misurazione disturba il sistema), bensì l'indeterminazione riflessiva del sistema che è il proprio oggetto. L'essere umano che riflette sui propri stati mentali non osserva quegli stati da fuori: li modifica attraverso l'osservazione. La metacognizione — la consapevolezza dei propri processi cognitivi — è un processo cognitivo che cambia i processi che osserva.

Questa proprietà del vivente — l'auto-referenzialità che produce indeterminazione strutturale — è il confine ultimo della neurobiologia. La neurobiologia può descrivere i meccanismi della metacognizione, può individuare i correlati neurali dell'auto-consapevolezza (la corteccia prefrontale mediale, la rete del default mode), può misurare come la consapevolezza metacognitiva modifica le decisioni e i giudizi. Ciò che non può fare — per la stessa ragione per cui Gödel ha dimostrato che un sistema formale non può dimostrare la propria consistenza dall'interno — è includere se stessa in quel circolo: la neurobiologia dell'auto-descrizione non può auto-descriversi senza entrare nel circolo riflessivo che è il proprio oggetto.

IX. Il confine come fondamento: perché i limiti non sono sconfitte

Un ciclo di ventiquattro monografie sulla biologia ai propri limiti non può concludersi con una resa. I limiti che questa monografia ha tracciato non sono sconfitte dell'impresa scientifica: sono il perimetro che definisce la forma dell'oggetto.

La geometria non è sconfitta perché non può misurare il profumo di una rosa. La neurobiologia non è sconfitta perché non può attraversare il divario esplicativo tra il processo fisico e l'esperienza soggettiva. Entrambe descrivono con precisione crescente le strutture dell'oggetto che studiano, e il confine di quella descrizione — il punto in cui il metodo raggiunge il suo limite — è tanto più informativo quanto più il metodo è stato rigoroso.

Il corpo vivente — il corpo come struttura bayesiana che minimizza l'energia libera, come sistema immunitario che risolve un gioco coevolutivo, come ippocampo che replay sequenze di eventi durante il sonno, come amigdala che computa il valore della minaccia, come corteccia che costruisce previsioni gerarchiche del mondo — è tutto questo. Ma è anche il corpo che tocca e viene toccato, che vede e viene visto, che comprende l'altro attraverso la risonanza e non attraverso l'inferenza, che si costituisce attraverso l'incontro con altri corpi in una struttura di intercorporeità irriducibile all'anatomia delle connessioni.

Il chiasmo di Merleau-Ponty — la reversibilità tra il toccante e il toccato che non si lascia catturare da nessuna descrizione in terza persona — non è una mistificazione poetica ma una struttura formale precisa: la struttura della prima persona, del punto di vista incarnato, che è la condizione di possibilità di ogni descrizione ma non coincide con nessuna descrizione. I neuroni specchio offrono il substrato biologico di quella struttura — il meccanismo attraverso cui il cervello risuona con i corpi altrui senza mediazione proposizionale — ma il meccanismo non è il chiasmo: è la sua realizzazione in terza persona, che lascia aperta la domanda sulla sua realizzazione in prima persona.

Questo è il confine. Non una lacuna da colmare ma una struttura da riconoscere. La biologia ai propri limiti non è la biologia in difetto: è la biologia nella propria forma più completa — quella in cui il metodo raggiunge il proprio bordo, si volge su se stesso, e riconosce che il corpo che studia è anche il corpo che studia, che l'oggetto della neurobiologia è anche il soggetto della neurobiologia, e che questa circolarità — lungi dall'essere un difetto logico — è la firma della cosa più interessante che la biologia possa studiare: un sistema che si auto-comprende, parzialmente, imperfettamente, strutturalmente incompletamente, e che in questa incompletezza strutturale non è meno meraviglioso, ma più.

Appendici tecniche

Appendice A — Monografia XVII: Il principio di energia libera variazionale e la stratificazione laminare del predictive coding

A.1 Derivazione dell'ELBO e della sua decomposizione

Sia $P(x, z)$ il modello generativo gerarchico e $q(z; \lambda)$ la distribuzione variazionale approssimante. Per la regola di Bayes:

$$P(z | x) = P(x, z) / P(x)$$

Poiché $P(x) = \int P(x, z) dz$ è generalmente intrattabile, si introduce la divergenza di Kullback-Leibler tra q e il posteriore esatto:

$$KL(q(z; \lambda) || P(z | x)) = E_q[\log q(z)] - E_q[\log P(z | x)]$$

Sostituendo $P(z | x) = P(x, z) / P(x)$:

$$KL(q || P) = E_q[\log q(z)] - E_q[\log P(x,z)] + \log P(x)$$

Poiché $KL \geq 0$:

$$\log P(x) \geq E_q[\log P(x,z)] - E_q[\log q(z)] \equiv ELBO(q)$$

L'energia libera variazionale è $F = -ELBO$. La decomposizione dell'ELBO in accuratezza meno complessità:

$$ELBO = E_q[\log P(x|z)] - KL(q(z) || P(z))$$

dove il primo termine misura la fedeltà del modello alle osservazioni (la log-verosimiglianza attesa) e il secondo misura la distanza dalla distribuzione a priori. Il cervello, secondo il principio di energia libera di Friston (2005), minimizza incessantemente F attraverso la variazione delle credenze q (inferenza) e dei parametri del modello P (apprendimento).

A.2 Derivazione della stratificazione laminare: frequenze gamma e beta nel predictive coding

Il modello laminare di Bastos et al. (2012) predice che i segnali di errore feedforward (strati II/III $\rightarrow$ strati IV dell'area superiore) si trasmettono in banda gamma, e le previsioni feedback (strati V/VI $\rightarrow$ strati I/II dell'area inferiore) si trasmettono in banda beta.

La giustificazione meccanicistica deriva dalla struttura delle sinapsi di feedback vs feedforward. Le connessioni feedforward dai neuroni piramidali degli strati II/III terminano prevalentemente su neuroni piramidali nello strato IV dell'area superiore, dove eccitano sia neuroni eccitatori sia interneuroni fast-spiking (FS) parvalbuminici che generano oscillazioni gamma locali attraverso il meccanismo PING (Pyramidal-Interneuron Network Gamma):

$$dE/dt = -E + F(w_{EE} \cdot E - w_{EI} \cdot I + \text{input}) \quad dI/dt = -I/\tau_I + F(w_{IE} \cdot E)$$

dove E è l'attività delle cellule eccitatorie, I degli interneuroni FS, τ_I è la costante di decadimento degli interneuroni (più piccola per i FS), e F è la funzione di attivazione non lineare. Il meccanismo PING produce oscillazioni gamma a $\omega_\gamma \approx 1/\tau_I$ (circa 40-80 Hz per i neuroni FS con $\tau_I \approx 5-10$ ms).

Le connessioni feedback dai neuroni piramidali degli strati V/VI terminano prevalentemente sugli strati I e II/III dell'area inferiore, dove eccitano neuroni piramidali con costanti di membrana più lente e interneuroni somatostatinergici (SST) con costanti di decadimento più lunghe:

$$\omega_{\beta} \approx 1/(\tau_{\{E\}} + \tau_{\{SST\}}) \approx 1/40-80 \text{ ms} \approx 12-25 \text{ Hz}$$

producendo oscillazioni in banda beta (13-30 Hz).

La predizione empirica è il *gradiente gamma/beta lungo la gerarchia visiva*: il rapporto gamma/beta aumenta progressivamente dalle aree inferiori a quelle superiori nelle proiezioni feedforward (le aree inferiori ricevono più feedback beta dagli strati superiori), e diminuisce nelle proiezioni feedback. Questo gradiente è stato documentato da Bastos et al. (2015) nella corteccia visiva del macaco con registrazioni LFP.

A.3 Derivazione formale della disregolazione della precisione nella psicosi

Sia Π_s la matrice di precisione degli errori di previsione sensoriali al livello 1 della gerarchia e Π_c la matrice di precisione delle previsioni al livello 2. L'aggiornamento della rappresentazione μ_2 al livello 2 è:

$$d\mu_2/dt = -\kappa[\Pi_s \cdot \epsilon_1 - f'_1(\mu_2)^T \cdot \Pi_c \cdot (\mu_2 - \mu_{2_prior})]$$

dove $\epsilon_1 = x_1 - f_1(\mu_2)$ è l'errore di previsione sensoriale e $(\mu_2 - \mu_{2_prior})$ è la deviazione dalla previsione a priori.

Per un sistema psicotico con $\Pi_s^{\{psicosi\}} = k_p \cdot \Pi_s^{\{normale\}}$ con $k_p \gg 1$ e $\Pi_c^{\{psicosi\}} = k_c \cdot \Pi_c^{\{normale\}}$ con $k_c \ll 1$:

$$d\mu_2/dt \approx -\kappa \cdot k_p \cdot \Pi_s^{\{normale\}} \cdot \epsilon_1$$

Il termine di regolarizzazione verso il prior diventa trascurabile ($k_c \ll 1$), e qualsiasi segnale di errore ϵ_1 — anche quelli generati da rumore sensoriale — produce un aggiornamento massiccio di μ_2 . Questo equivale all'interpretazione causale di segnali irrilevanti — la struttura formale dell'aberrant salience (Kapur, 2003).

La correzione farmacologica con antagonisti D2 opera riducendo k_p verso 1 (ripristinando la normale precisione degli errori sensoriali) attraverso la riduzione del neuromodulatore dopaminergico che nella corteccia prefrontale potenzia la guadagno dei neuroni piramidali della via feedforward, aumentando effettivamente la precisione dei segnali di errore.

Appendice B — Monografia XVIII: Derivazione della struttura concettuale CES e convergenza del PCI

B.1 Costruzione della MIP e derivazione della phi del meccanismo

Sia M un sistema con stati discreti, $s \subseteq M$ un meccanismo, e s_c il suo stato corrente. La distribuzione causa-effetto congiunta è:

$$\pi^{\{ce\}}(s \rightarrow M) = \pi^c(s \rightarrow M) \times \pi^e(s \rightarrow M)$$

dove $\times$ indica il prodotto cartesiano delle due distribuzioni marginali. La partizione di un meccanismo s in (s_1, s_2) con $s_1 \cup s_2 = s$ e $s_1 \cap s_2 = \emptyset$ produce:

$$\pi^{\{ce\}}_{\{partition\}}(s_1, s_2) = \pi^{\{ce\}}(s_1 \rightarrow M_1) \times \pi^{\{ce\}}(s_2 \rightarrow M_2)$$

dove M_1 e M_2 sono le parti del sistema che interagiscono con s_1 e s_2 rispettivamente. La MIP è la partizione che minimizza la perdita di informazione causale:

$$MIP(s) = \underset{(s_1, s_2)}{\operatorname{argmin}} \{ EMD(\pi^{\{ce\}}(s \rightarrow M), \pi^{\{ce\}}_{\{partition\}}(s_1, s_2)) \}$$

La ϕ del meccanismo è:

$$\phi(s) = EMD(\pi^{\{ce\}}(s \rightarrow M), \pi^{\{ce\}}_{\{MIP(s)\}}(s \rightarrow M))$$

Proprietà fondamentale: $\phi(s) = 0$ per qualsiasi meccanismo s che può essere completamente fattorizzato, e $\phi(s) > 0$ soltanto per meccanismi con dipendenze causali irriducibili tra le proprie parti.

B.2 Derivazione del PCI e della sua soglia diagnostica

Il PCI (Perturbational Complexity Index, Casali et al. 2013) è:

$$PCI = C_{\{LZ\}}(\text{matrice_risposta_TMS_binarizzata}) / \log_2(n_t \cdot n_e)$$

dove $C_{\{LZ\}}$ è la complessità di Lempel-Ziv della matrice binarizzata al segno e il denominatore è la normalizzazione per la dimensione massima possibile della matrice (n_t campioni temporali $\times$ n_e elettrodi).

Derivazione della soglia di 0.31. Casarotto et al. (2016) hanno calibrato la soglia su 150 pazienti in stati coscienti e non coscienti verificati. La distribuzione del PCI in stati non coscienti segue una distribuzione Gaussiana con $\mu \approx 0.22$, $\sigma \approx 0.05$; in stati coscienti, $\mu \approx 0.51$, $\sigma \approx 0.08$. La soglia ottimale di separazione secondo il criterio di Youden (massimizzazione di sensibilità + specificità - 1) cade a $PCI^* = 0.31$, con sensibilità ≈ 0.97 e specificità ≈ 0.95 .

Appendice C — Monografia XIX: Potere generativo delle grammatiche CF e neurobiologia di FOXP2

C.1 Dimostrazione del pumping lemma per grammatiche context-free

Teorema (Pumping Lemma per CF): Se L è un linguaggio context-free, esiste un intero p (pumping length) tale che per qualsiasi stringa $w \in L$ con $|w| \geq p$, w può essere scritta come $w = uvxyz$ con:

1. $|vy| \geq 1$
2. $|vxy| \leq p$
3. Per ogni $i \geq 0$, $uv^i xy^i z \in L$

Corollario: Il linguaggio $\{a^n b^n c^n : n \geq 1\}$ non è context-free (per pumping argument). Le lingue naturali con struttura $a^n b^n c^n$ richiedono almeno grammatiche context-sensitive.

Il linguaggio delle dipendenze incrociate nel tedesco e nel ceco (documentato da Shieber, 1985) ha struttura formale non context-free — prova che il linguaggio naturale supera la potenza delle grammatiche CF e richiede la struttura delle grammatiche mildly context-sensitive (TAG, HPSG).

C.2 Derivazione formale dell'effetto FOXP2 sui gangli della base

L'ipotesi FOXP2-ganglia-basale-sequenza predice che la versione umana di FOXP2 regola l'espressione di geni nel circuito cortico-striato-talamo-corticale che implementa la selezione di sequenze motorie gerarchicamente strutturate.

Sia S una sequenza motoria di livello gerarchico k , e sia $V(S, n)$ il valore atteso della sequenza S quando sono stati già prodotti n elementi. La selezione della sequenza è:

$$a_{\{n+1\}} = \operatorname{argmax}_{\{a\}} Q(S + a, V(S, n), \theta_{\{\text{FOXP2}\}})$$

dove $\theta_{\{\text{FOXP2}\}}$ sono i parametri del circuito striatale che dipendono dall'espressione di FOXP2. La versione umana di FOXP2 (con T303N e N325S) regola positivamente CNTNAP2, DISC1 e altri geni della plasticità corticostriatale, aumentando effettivamente la capacità del circuito di rappresentare sequenze di livello gerarchico $k > 2$ — il che corrisponde computazionalmente all'estensione del potere generativo del sistema da grammatiche regolari a grammatiche context-free.

Appendice D — Monografia XX: Derivazione del synaptic tagging and capture e delle oscillazioni theta come gate temporale

D.1 Modello formale del synaptic tagging and capture

Sia $\text{tag}(s, t) \in \{0, 1\}$ lo stato del tag sinaptico alla sinapsi s al tempo t , $\text{PRP_conc}(t)$ la concentrazione di proteine plasticità-correlate nel dendrita, e $w(s, t)$ il peso sinaptico.

La dinamica del tag:

$\text{tag}(s, t) = 1$ se $e_{\text{induz}}(s, t) > \theta_{\text{induction}}$ e $t \in [t_{\text{LTP}} - \tau_{\text{pre}}, t_{\text{LTP}}]$ $\text{tag}(s, t) \rightarrow 0$ per $t > t_{\text{LTP}} + \tau_{\text{tag}}$ con $\tau_{\text{tag}} \approx 1-3$ ore

La concentrazione di PRP:

$$\frac{d\text{PRP_conc}}{dt} = -\text{PRP_conc}/\tau_{\{\text{PRP}\}} + \sum_{s'} \sigma(w(s', t) - w(s', t-1)) \cdot 1_{\{\text{tag}(s', t)=1\}}$$

dove il termine di produzione è proporzionale all'induction di LTP in qualsiasi sinapsi del dendrita (produzione di PRP nel soma per attivazione PKA-CREB).

La regola di consolidazione L-LTP:

$$w(s, t) \rightarrow w_{\{\text{L-LTP}\}}(s) \text{ se } \text{tag}(s, t) > 0 \text{ e } \text{PRP_conc}(t) > \theta_{\{\text{PRP}\}}$$

Predizione cruciale del STC: la stimolazione debole (E-LTP ma non L-LTP) in una sinapsi s_1 , seguita da stimolazione forte (L-LTP) in una sinapsi s_2 dello stesso neurone entro τ_{tag} , produce L-LTP anche in s_1 — perché le PRP prodotte per s_2 vengono catturate da $\text{tag}(s_1)$. Questa predizione è stata verificata sperimentalmente da Frey e Morris (1997).

D.2 Gate oscillatorio theta per la separazione encoding/retrieval

Le oscillazioni theta (4-8 Hz) separano temporalmente le fasi di encoding (fase 270°) e retrieval (fase 90°). La regola di apprendimento con gate theta è:

$$\Delta w_{\{ij\}}(t) = \eta \cdot x_i(t) \cdot x_j(t) \cdot G_{\theta}(t)$$

dove:

$G_{\theta}(t) = 1$ se $\varphi_{\theta}(t) \in [\theta_{\{\text{enc}, \text{min}\}}, \theta_{\{\text{enc}, \text{max}\}}]$ (finestra di encoding) $G_{\theta}(t) = 0$ se $\varphi_{\theta}(t) \in [\theta_{\{\text{ret}, \text{min}\}}, \theta_{\{\text{ret}, \text{max}\}}]$ (finestra di retrieval)

con $\varphi_{\theta}(t)$ la fase istantanea dell'oscillazione theta. Il modello predice che:

1. La LTP può essere indotta soltanto durante la finestra di encoding (fase 270°)
2. Il recupero dalle sinapsi CA3→CA1 avviene massimalmente durante la fase di retrieval (fase 90°)
3. La separazione delle due finestre previene l'interferenza tra encoding e retrieval

Derivazione della coerenza con il meccanismo colinergico. La modulazione colinergica (acetilcolina, rilasciata dal nucleo basale di Meynert durante l'esplorazione attiva) sopprime le connessioni ricorrenti CA3→CA1 e potenzia la via EC→CA1. La soppressione di CA3→CA1 ha costante di tempo $\tau_{\{ACh\}} \approx 50-100$ ms, corrispondente alla semi-lunghezza d'onda del ciclo theta — il che garantisce automaticamente che la soppressione del retrieval (via CA3) coincida con il picco dell'encoding (via EC).

Appendice E — Monografia XXI: Derivazione del potenziale affettivo $U(v,a)$ e stima dei parametri

E.1 Derivazione del sistema dinamico affettivo dai dati EMA

Sia $v(t)$ la valenza e $a(t)$ l'arousal misurati in serie temporali EMA con intervallo $\Delta t = 30$ min. Il modello stocastico continuo è:

$$dv/dt = -\gamma_v(v - v_{SPA}) + \sum_k \xi_k(t) \cdot P_k + \sigma_v \cdot W_v(t)$$

dove $W_v(t)$ è il processo di Wiener standard. La versione discreta per la stima dai dati è:

$$v_{\{t+1\}} = v_t - \gamma_v \Delta t \cdot (v_t - v_{SPA}) + \sum_k \xi_k \cdot P_k \Delta t + \sigma_v \sqrt{\Delta t} \cdot \epsilon_t$$

con $\epsilon_t \sim N(0,1)$. Questo è un modello AR(1) con tendenza verso la media:

$$v_{\{t+1\}} = (1 - \gamma_v \Delta t) \cdot v_t + \gamma_v \Delta t \cdot v_{SPA} + \text{disturbi}$$

Il coefficiente di autocorrelazione di lag 1 è:

$$\rho_1 = \text{Corr}(v_{\{t+1\}}, v_t) = 1 - \gamma_v \Delta t$$

Da $\rho_1 \approx 0.40-0.60$ (Kuppens et al., 2010) con $\Delta t = 30$ min = 0.5 h:

$$\gamma_v \approx (1 - q_1) / \Delta t \approx 0.40-0.60 / 0.5 \text{ ore} = 0.80-1.20 / \text{ora}$$

Correggendo per la misurazione discreta (correzione di Pearson per AR(1)):

$$\gamma_v^{\{\text{corretto}\}} \approx 0.40-0.65 / \text{ora per alta regolazione}$$

$$\gamma_v^{\{\text{depressione}\}} \approx 0.10-0.25 / \text{ora}$$

Il potenziale termodinamico equivalente. L'energia libera del sistema affettivo (in assenza di disturbi stocastici e segnali esterni) è:

$$U(v, a) = (\gamma_v / 2)(v - v_{\text{SPA}})^2 + (\gamma_a / 2)(a - a_{\text{SPA}})^2$$

La forza di richiamo è proporzionale alla pendenza del potenziale: $F = -\partial U / \partial v = -\gamma_v(v - v_{\text{SPA}})$, coerente con la dinamica di richiamo verso lo SPA.

Appendice F — Monografia XXII: Dinamica Lotka-Volterra CTL-virus e coevoluzione dei quasispecies

F.1 Analisi dell'equilibrio e della stabilità del sistema CTL-virus

Il sistema di quattro equazioni ha più equilibri. L'equilibrio di infezione cronica con risposta CTL è:

$$I = d_E / q^* E = (p/c \cdot \beta \cdot T - \delta) / k^{**} \quad V = p \cdot I / c^{**} \quad T = \lambda / (d_T + \beta \cdot V)^{**}$$

La stabilità locale attorno all'equilibrio si analizza linearizzando il sistema. La matrice Jacobiana J all'equilibrio di infezione cronica è:

$$J = \begin{bmatrix} [-d_T - \beta \cdot V, 0, -\beta \cdot T, 0],^{**} &^{**} & [\beta \cdot V^*, -\delta - k \cdot E^*, \beta \cdot T^*, -k \cdot I^*],^{**} \\^{**} &^{**} & [0, p, -c, 0],^{**} &^{**} & [0, q \cdot E^*, 0, q \cdot I^* - d_E]^{**} \end{bmatrix}^{**}$$

Con $d_E = q \cdot I^*$ all'equilibrio, la traccia della jacobiana è $\text{Tr}(J) = -d_T - \beta \cdot V^* - \delta - k \cdot E^* - c < 0$ (tutti i termini diagonali sono negativi). La condizione per oscillazioni smorzate (autovalori complessi con parte immaginaria non nulla) è:

$$\text{Im}(\lambda) \neq 0 \leftrightarrow (\text{Tr}(J)/2)^2 < \det(J)$$

che è soddisfatta per un ampio intervallo di parametri fisio-logicamente realistici, producendo le oscillazioni smorzate osservate nella fase acuta delle infezioni virali.

F.2 Derivazione della soglia Red Queen per il polimorfismo HLA

La condizione per il mantenimento del polimorfismo con selezione dipendente dalla frequenza negativa è:

$$dw_i/dp_i < 0 \text{ per tutti gli alleli } i$$

ossia la fitness di un allele deve diminuire quando la sua frequenza aumenta. Per il sistema HLA-patogeno, la fitness dell'aplotipo HLA_i è:

$$w_i(p) = w_0 - \alpha \cdot p_i \cdot \beta_{\{escape\}}(p_i, t)$$

dove $\beta_{\{escape\}}(p_i, t)$ è la proporzione di patogeni che, al tempo t , si sono adattati all'aplotipo HLA_i con alta frequenza p_i . Poiché $\beta_{\{escape\}}$ aumenta con p_i (i patogeni evolvono per sfuggire agli aplotipi più comuni), $dw_i/dp_i < 0$ e il polimorfismo è mantenuto.

La condizione di equilibrio polimorfistico è:

$$p_i = (w_0 - \bar{w}) / (a \cdot \beta_{\{escape\}}'(p_i))^{**}$$

dove $\bar{w}$ è la fitness media e $\beta_{\{escape\}}'$ è la derivata di $\beta_{\{escape\}}$ rispetto a p_i . Poiché questa equazione ha soluzioni per tutti gli alleli $i = 1, \dots, N$, il polimorfismo è mantenuto in tutto il pool allelico — il che spiega i >30.000 alleli HLA osservati.

Appendice G — Monografia XXIII: Derivazione della cinetica UPR e formalizzazione della biologia dei telomeri

G.1 Modello cinetico della decisione UPR: adattamento versus apoptosi

Sia $u(t)$ l'attività di IRE1 α , $p(t)$ l'attività di PERK (misurata come livello di fosfo-eIF2 α) e $c(t)$ il livello di CHOP. Il modello cinetico semplificato è:

$$\begin{aligned} du/dt &= k_u \cdot S(t) / (K_u + S(t)) - \delta_u \cdot u \\ dp/dt &= k_p \cdot S(t) / (K_p + S(t)) - \delta_p \cdot p \\ dc/dt &= k_c \cdot p / (K_c + p) - \delta_c \cdot c \end{aligned}$$

dove $S(t)$ è il livello di proteine mal-ripiegate nel RE.

La decisione adattamento vs. apoptosi dipende dalla cinetica di $S(t)$:

Stress transitorio: $S(t) = S_{\max}$ per $t \in [0, \tau_{\text{stress}}]$ con $\tau_{\text{stress}} < \tau_{\text{adapt}}$. In questo caso u e p raggiungono valori massimi ma l'integrale

$\int c(t)dt$ rimane sotto la soglia apoptotica θ_{apopt} perché CHOP non ha tempo di accumularsi:

$$\int c(t)dt < \theta_{\text{apopt}} \rightarrow \text{adattamento}$$

Stress prolungato: $S(t) = S_{\text{max}}$ per $t > \tau_{\text{stress}}$. In questo caso $c(t)$ si accumula progressivamente e supera θ_{apopt} :

$$\int c(t)dt > \theta_{\text{apopt}} \rightarrow \text{apoptosi}$$

La dipendenza dalla cinetica (non dalla sola ampiezza) è la firma formale della distinzione fisiologica tra stress acuto adattativo e stress cronico patologico.

G.2 Formalizzazione del limite di Hayflick e del SASP

La lunghezza del telomero più corto al ciclo n :

$$L_n = L_0 - \Delta l \cdot n - \sigma_L \cdot \sqrt{n} \cdot Z_n$$

dove il termine stocastico $\sigma_L \cdot \sqrt{n} \cdot Z_n$ (con $Z_n \sim N(0,1)$) riflette la variabilità dell'accorciamento tra divisioni. La senescenza viene innescata quando $L_n < L_{\text{cr}} = 5 \text{ kb}$. Il numero di divisioni attese fino alla senescenza è:

$$N_{\text{sens}} = (L_0 - L_{\text{cr}}) / \Delta l \approx 100 \text{ divisioni (stima deterministica)}$$

Con variabilità stocastica, la distribuzione di N_{sens} è approssimativamente gaussiana con media $\mu \approx 100$ e deviazione standard $\sigma \approx \sqrt{(\sigma_L^2 \cdot \mu) / \Delta l^2}$.

Il SASP come sistema di feedback: sia $q_{\text{SASP}}(t)$ la densità di cellule senescenti nel tessuto. Il SASP produce citochine pro-infiammatorie che inducono senescenza nelle cellule adiacenti (senescenza paracrina) con tasso k_{SASP} , e attivano le cellule immunitarie per la clearance delle cellule senescenti con tasso k_{clear} . Il bilancio è:

$$dq_{\text{SASP}}/dt = r_{\text{div}} \cdot (1 - q_{\text{SASP}}) \cdot f(L) + k_{\text{SASP}} \cdot q_{\text{SASP}} - k_{\text{clear}} \cdot q_{\text{SASP}} \cdot I(t)$$

dove r_{div} è il tasso di divisione del tessuto, $f(L)$ è la probabilità di entrare in senescenza per lunghezza media dei telomeri L , e $I(t)$ è l'intensità della risposta immunitaria. In assenza di adeguata risposta immunitaria (come nell'invecchiamento), q_{SASP} cresce a causa del termine SASP autocatalitico $k_{\text{SASP}} \cdot q_{\text{SASP}}$ — un processo di feedback positivo che accelera l'invecchiamento tissutale.

Appendice H — Monografia XXIV: Struttura formale del chiasmo e dell'*embodied simulation*

H.1 Il chiasmo come struttura di punto fisso del corpo

Sia B lo spazio dei corpi (fisici) e E lo spazio delle esperienze corporee (fenomenologiche). Il chiasmo di Merleau-Ponty può essere formalizzato come la struttura di una coppia di funzioni:

$\varphi: B \rightarrow E$ (la percezione: il corpo fisico produce esperienza) $\psi: E \rightarrow B$ (l'azione: l'esperienza produce comportamento corporeo)

La circolarità chiasmatica è la struttura $\psi \circ \varphi: B \rightarrow B$ — il ciclo sensomotorio che connette l'input corporeo al output corporeo attraverso l'esperienza. Per il teorema di Brouwer sul punto fisso, se $\psi \circ \varphi$ è continua e B è compatto e convesso, esiste almeno un punto fisso $b^* \in B$ tale che $\psi(\varphi(b^*)) = b^*$. Questo punto fisso è lo *schema corporeo di equilibrio*: lo stato del corpo fisico che si riproduce attraverso il ciclo sensomotorio dell'esperienza.

L'incompletezza della reversibilità chiasmatica corrisponde formalmente al fatto che φ e ψ non sono inverse esatte l'una dell'altra: $\varphi \circ \psi \neq \text{id}_E$. L'esperienza del corpo dell'altro non è mai identica all'esperienza del proprio corpo — c'è sempre un residuo irriducibile che è la prima persona dell'altro, non accessibile dalla propria esperienza corporea.

H.2 L'*embodied simulation* come inferenza analogica bayesiana

La *embodied simulation* può essere formalizzata come un'inferenza bayesiana il cui spazio delle ipotesi è lo spazio degli stati motori/emotivi propri del soggetto S . Dato l'osservazione del movimento/espressione dell'individuo O (l'evidenza), il sistema nervoso di S inferisce lo stato interno di O attraverso il prior sull'autospazio:

$$P(\text{stato}_O \mid \text{osservazione}_O) \propto P(\text{osservazione}_O \mid \text{stato}_S) \cdot P(\text{stato}_S)$$

dove il likelihood $P(\text{osservazione}_O \mid \text{stato}_S)$ è calcolato simulando l'output osservabile che corrisponderebbe allo stato S nel proprio sistema motorio/emotivo. La likelihood è alta quando stato_S produce, simulato, un output vicino all'osservazione_ O .

Questo meccanismo predice:

1. La comprensione delle azioni/emozioni altrui è mediata dall'attivazione del proprio sistema motorio/emotivo — la firma del sistema specchio.

2. La comprensione è più accurata per azioni/emozioni che il soggetto è in grado di eseguire/provare — predizione verificata sperimentalmente per le emozioni (maggiore accuratezza del riconoscimento per le emozioni che il soggetto prova abitualmente).
3. La comprensione degrada quando la simulazione è resa difficile dalla disconnessione motoria — predizione verificata con la paralisi facciale indotta da iniezione di botulino che riduce l'accuratezza del riconoscimento delle espressioni facciali (Niedenthal et al., 2009).

Appendice I — Tabelle di corrispondenza formale tra i tre movimenti

I.1 Isomorfismi strutturali tra i cicli

Struttura formale	Ai confini del calcolabile	Ai confini del vivente
Teorema di impossibilità fondamentale	Gödel-Turing (incompletezza)	Teorema di Rice biologico (invecchiamento irriducibile)
Principio unificante	ELBO / energia libera di Helmholtz	ELBO / energia libera variazionale di Friston
Struttura di gioco coevolutivo	Giochi evolutivi cognitivi (MARL)	Red Queen HLA-patogeni
Limite fenomenologico	Indeterminazione radicale (Quine-Davidson)	Chiasmo (Merleau-Ponty)
Sistema di error correction	Verifica formale parziale (Reluplex)	DNA repair + UPR + autofagia + telomeri
Fondamento evolutivo	Algoritmi evolutivi, NK model	Equazione di Price, selezione naturale
Inferenza ottimale	AIXI / prior di Solomonoff	Cervello bayesiano / prior biologico
Struttura semantica	Semantica formale, grounding	Semantica evolutiva, catena kripkiana biologica

I.2 La gerarchia delle precisioni nel predictive coding e la psicopatologia

Disturbo	Π_s (errori sensoriali)	Π_c (previsioni)	γ_v (regolazione affettiva)	Meccanismo principale
Stato normale	Π_s^0	Π_c^0	$\gamma_v \approx 0.5/\text{ora}$	Bilanciamento ottimale
Psicosi	$\Pi_s \gg \Pi_s^0$	$\Pi_c \ll \Pi_c^0$	variabile	Iperprecisione errori sensoriali (DA)
Ansia	variabile	$\Pi_{\text{minaccia}} \gg \Pi_{\text{minaccia}}^0$	γ_v ridotto	Iperprecisione previsioni di pericolo
Depressione	$\Pi_{\text{ricompensa}} \ll \Pi_{\text{ricompensa}}^0$	$\Pi_{\text{negativa}} \gg \Pi_{\text{negativa}}^0$	$\gamma_v \approx 0.15/\text{ora}$	Ipoprecisione segnali di rinforzo
DBP	variabile (alta reattività)	variabile	$\gamma_v \approx 0.8/\text{ora}$ ma σ_v alta	Alta reattività con alta rumorosità

I.3 Notazione unificata

Simboli bayesiani e variazionali:

- $P(H)$, $P(H|D)$, $P(D|H)$: probabilità a priori, a posteriori, verosimiglianza
- $q(z;\lambda)$: distribuzione variazionale approssimante parametrizzata da λ
- $F = -\text{ELBO}$: energia libera variazionale
- $\text{KL}(q||P)$: divergenza di Kullback-Leibler
- $\text{ELBO} = E_q[\log P(x,z)] - E_q[\log q(z)]$

Simboli del predictive coding:

- ϵ_l : vettore di errore di previsione al livello l
- μ_l : vettore di rappresentazione al livello l
- Σ_l : matrice di covarianza del rumore al livello l
- $\Pi_l = \Sigma_l^{-1}$: matrice di precisione al livello l
- f_l : funzione di generazione dal livello $l+1$ al livello l

Simboli neurobiologici della memoria:

- w_{ij} : peso sinaptico dalla sinapsi pre i alla post j
- $\text{tag}(s,t)$: stato del tag sinaptico (0 o 1) alla sinapsi s al tempo t
- $\text{PRP}_{\text{conc}}(t)$: concentrazione di proteine plasticità-correlate
- $\varphi_\theta(t)$: fase istantanea delle oscillazioni theta

Simboli affettivi:

- $v(t)$: valenza al tempo t
- $a(t)$: arousal al tempo t
- $v_{\text{SPA}}, a_{\text{SPA}}$: set point affettivo (valenza e arousal basali)
- γ_v, γ_a : costanti di smorzamento affettivo
- $U(v,a)$: potenziale termodinamico affettivo

Simboli evolutivi e immunologici:

- $w_i(p)$: fitness dell'aplotipo i come funzione del vettore delle frequenze p
- p_i : frequenza dell'aplotipo HLA i
- $T(t), I(t), V(t), E(t)$: cellule target, infettate, virioni liberi, CTL al tempo t
- ϵ_j : sensibilità del ceppo virale j alla risposta CTL
- $\mu_{\{kj\}}$: tasso di mutazione del ceppo k verso il ceppo j

Simboli telomerici:

- L_n : lunghezza del telomero più corto dopo n divisioni
- L_0 : lunghezza iniziale dei telomeri
- Δ : accorciamento per divisione (≈ 100 bp)
- L_{cr} : lunghezza critica minima (≈ 5 kb)
- n_{max} : numero massimo di divisioni (≈ 100)

Appendice L — Indice dei teoremi e proposizioni del terzo movimento

Monografia XVII — La mente bayesiana

Teorema XVII.1 (Convergenza di AIXItl a AIXI). Per qualsiasi sequenza di osservazioni $x_1 x_2 \dots x_n$, $\lim_{\{l, t \rightarrow \infty\}} M_{\{tl\}}(x_1 \dots x_n) = M(x_1 \dots x_n)$ dove M è la misura universale di Solomonoff. → *Sezione VII.1, Appendice A.3*

Proposizione XVII.2 (Disregolazione della precisione nella psicosi). Per $\Pi_s^{\{\text{psicosi}\}} = k_p \cdot \Pi_s^{\{\text{normale}\}}$ con $k_p \gg 1$ e $\Pi_c^{\{\text{psicosi}\}} = k_c \cdot \Pi_c^{\{\text{normale}\}}$ con $k_c \ll 1$, la dinamica di aggiornamento delle rappresentazioni produce interpretazione causale di segnali di rumore (aberrant salience). → *Sezione VI, Appendice A.3*

Proposizione XVII.3 (Gradiente gamma/beta nella gerarchia corticale). Il rapporto $\text{potenza_gamma}/\text{potenza_beta}$ nella coerenza LFP delle proiezioni feedforward è maggiore di 1 per le connessioni ascendenti nella gerarchia visiva, e inferiore a 1 per le connessioni discendenti, con un gradiente monotono dalla V1 alla PFC. → *Sezione IV, Appendice A.2*

Monografia XVIII — La neurobiologia della coscienza

Teorema XVIII.1 (Proprietà di irreducibilità della CES). Per qualsiasi sistema M con $\Phi(M) > 0$, la struttura concettuale $CES(M)$ non è il prodotto cartesiano delle CES dei sottosistemi propri di M ; ossia non esiste una partizione P di M tale che $CES(M) = CES(M_1) \times CES(M_2)$. → *Sezione II.3*

Proposizione XVIII.2 (Soglia diagnostica del PCI). La soglia $PCI^* = 0.31$ separa stati coscienti e non coscienti con sensibilità ≈ 0.97 e specificità ≈ 0.95 in una popolazione di 150 pazienti validati. → *Sezione II.5, Appendice B.2*

Proposizione XVIII.3 (Localizzazione posteriore degli NCC in COGITATE). Le registrazioni sEEG, MEG e fMRI nell'adversarial collaboration di Melloni et al. (2023) mostrano che l'attivazione della corteccia temporale posteriore discrimina il contenuto cosciente per l'intera durata dell'esperienza (100-1500 ms), mentre l'attivazione prefrontale discrimina principalmente il report (> 400 ms). → *Sezione V*

Monografia XIX — La semantica evolutiva

Teorema XIX.1 (Superiorità generativa delle CF sulle grammatiche regolari). Per qualsiasi grammatica regolare G_R , esiste un linguaggio L_{CF} generato da una grammatica context-free che non può essere generato da G_R . Il linguaggio $\{a^n b^n : n \geq 1\}$ è l'esempio canonico. → *Sezione V.1, Appendice C.1*

Proposizione XIX.2 (Effetto FOXP2 sull'apprendimento procedurale). I topi con le sostituzioni umane T303N e N325S in FOXP2 mostrano: (a) vocalizzazioni ultrasoni più variate; (b) plasticità sinaptica striatale aumentata; (c) acquisizione più rapida di sequenze motorie strutturate. → *Sezione V.2, Appendice C.2*

Proposizione XIX.3 (Composizionalità di Montague). Per qualsiasi espressione complessa $E = E_1 * E_2$ dove $*$ è una operazione sintattica di composizione, il significato di E è determinato composizionalmente: $\llbracket E \rrbracket = \llbracket E_1 \rrbracket * \llbracket E_2 \rrbracket$ con $*$ implementata come applicazione funzionale sui tipi semantici. → *Sezione VII.2*

Monografia XX — La struttura formale della memoria

Teorema XX.1 (Convergenza del tag sinaptico). Nelle condizioni del modello STC (Frey-Morris, 1997), una sinapsi con tag attivo cattura le PRP se e solo se la concentrazione $PRP_{conc}(t)$ supera la soglia $\theta_{\{PRP\}}$ durante la finestra di labilità τ_{tag} del tag. La probabilità di cattura decresce esponenzialmente con la distanza temporale tra induction del tag e picco di PRP: $P_{capture} \propto \exp(-\Delta t/\tau_{tag})$. → *Sezione III, Appendice D.1*

Proposizione XX.2 (Separazione di fase encoding/retrieval nelle oscillazioni theta). Per un ciclo theta con periodo $T_{\theta} \approx 125-250$ ms, la separazione della finestra di encoding ($\varphi \approx 270^\circ$) dalla finestra di retrieval ($\varphi \approx 90^\circ$) è di $T_{\theta}/2 \approx 62-125$ ms — sufficientemente lunga da impedire la sovrapposizione dei processi hebbiani di potenziamento e di soppressione. → *Sezione IV, Appendice D.2*

Proposizione XX.3 (Compressione temporale del replay SWR). La sequenza di attivazione delle cellule luogo durante l'esplorazione di durata $\tau_{exp} \approx 1-10$ minuti viene compressa in un replay durante un SWR di durata $\tau_{SWR} \approx 50-100$ ms, con fattore di compressione $C = \tau_{exp}/\tau_{SWR} \approx 6000x$ (Wilson e McNaughton, 1994). → *Sezione V.2*

Teorema XX.4 (Labilità selettiva nella reconsolidazione). Il blocco della sintesi proteica nell'amigdala basolaterale immediatamente dopo il retrieval di un ricordo di paura condizionata abolisce il ricordo nelle prove successive, senza effetto se il blocco avviene prima del retrieval o più di 6 ore dopo (Nader, Schafe e LeDoux, 2000). La finestra di labilità è di 4-6 ore dalla fine del retrieval. → *Sezione VI.1*

Monografia XXI — La fisica dell'emozione

Proposizione XXI.1 (Forma del potenziale affettivo $U(v,a)$). La dinamica degli stati affettivi nella popolazione di Kuppens et al. (2010) è coerente con un potenziale parabolico $U(v,a) = (\gamma_v/2)(v-v_{SPA})^2 + (\gamma_a/2)(a-a_{SPA})^2$ con $\gamma_v \approx 0.40-0.65/\text{ora}$ per individui con alta regolazione affettiva e $\gamma_v \approx 0.10-0.25/\text{ora}$ per individui con bassa regolazione (inclusa la depressione maggiore). → *Sezione VI, Appendice E.1*

Proposizione XXI.2 (Efficacia terapeutica degli SSRI come modulatori di γ_v). Dopo 4-8 settimane di trattamento con SSRI, la costante di smorzamento γ_v aumenta significativamente rispetto alla baseline in pazienti con depressione maggiore — corrispondente alla riduzione dell'inertia affettiva documentata clinicamente. → *Sezione VII.1*

Proposizione XXI.3 (Via rapida e lenta dell'amigdala). La via rapida talamo→BLA (latenza 12 ms) trasmette informazioni grossolane di minaccia e può attivare la risposta di paura prima dell'elaborazione corticale. La via lenta talamo→corteccia→BLA (latenza 25-50 ms) trasmette informazioni più elaborate e media la modulazione cognitiva della risposta emotiva. → *Sezione V.3*

Monografia XXII — I giochi evolutivi della neurobiologia

Teorema XXII.1 (Mantenimento del polimorfismo HLA per selezione dipendente dalla frequenza). Se la fitness $w_i(p)$ dell'aplotipo HLA_i è una funzione strettamente decrescente di p_i (selezione dipendente dalla frequenza negativa), allora il sistema evolutivo ha un equilibrio polimorfistico stabile con $p_i^* > 0$ per tutti gli alleli i . → *Sezione II.2, Appendice F.2*

Proposizione XXII.2 (Oscillazioni del sistema CTL-virus). Il sistema Lotka-Volterra CTL-virus produce oscillazioni smorzate attorno all'equilibrio di infezione cronica con frequenza $\omega \approx \sqrt{(k \cdot q \cdot I^* \cdot E^* - (\delta + k \cdot E^*)^2)/4}$. Per parametri HIV tipici, $\omega \approx 0.1\text{-}0.5/\text{settimana}$. → *Sezione II.3, Appendice F.1*

Proposizione XXII.3 (Numero di Dunbar dalla regressione neocorticale). Per la regressione $\log(N) = a + b \cdot \log(NQ)$ su 38 specie di primati con $b \approx 3.8$, l'applicazione al quoziente neocorticale umano $NQ \approx 4.1$ produce $N_{\text{Dunbar}} \approx 148$. → *Sezione IV.1*

Monografia XXIII — La verifica biologica

Teorema XXIII.1 (Accumulo irriducibile di errori nell'invecchiamento). Per qualsiasi sistema di verifica biologica con k meccanismi di rilevamento a precisione $p < 1$, la probabilità che un errore sfugga a tutti i k meccanismi è $P_{\text{err}} = p^k$. Poiché p e k sono finiti e i meccanismi di riparazione si degradano nel tempo, l'accumulo netto di danni cresce monotonicamente — producendo l'invecchiamento come processo irriducibile. → *Sezione VI*

Proposizione XXIII.2 (Decisione UPR dipendente dalla cinetica). Lo stress del RE di uguale ampiezza massima produce adattamento se transitorio ($\int c(t)dt < \theta_{\text{apopt}}$) e apoptosi se prolungato ($\int c(t)dt > \theta_{\text{apopt}}$). La distinzione è determinata dalla cinetica del segnale, non dall'ampiezza istantanea. → *Sezione III.3, Appendice G.1*

Proposizione XXIII.3 (Formula del limite di Hayflick). La lunghezza del telomero più corto dopo n divisioni è $L_n = L_0 - \Delta l \cdot n$ con $L_0 \approx 15 \text{ kb}$ e $\Delta l \approx 100 \text{ bp}$. La senescenza viene innescata quando

$L_n < L_{cr} \approx 5$ kb, producendo $n_{max} \approx 100$ divisioni teoriche (limite di Hayflick osservato ≈ 50 per accorciamento in vivo precedente la coltura). → *Sezione V.2, Appendice G.2*

Monografia XXIV — Il corpo come indeterminazione

Proposizione XXIV.1 (Punto fisso dello schema corporeo nel chiasmo). La circolarità sensomotoria $\psi \circ \varphi: B \rightarrow B$ ha almeno un punto fisso b^* per il teorema di Brouwer (B compatto e convesso, $\psi \circ \varphi$ continua). Questo punto fisso corrisponde allo schema corporeo di equilibrio — lo stato del corpo che si riproduce attraverso il ciclo sensomotorio dell'esperienza. → *Sezione III, Appendice H.1*

Proposizione XXIV.2 (Embodied simulation come inferenza bayesiana). La comprensione delle azioni altrui nel sistema specchio è formalmente equivalente a un'inferenza bayesiana con likelihood $P(\text{osservazione}_O \mid \text{stato}_S)$ calcolato simulando l'output del proprio sistema motorio nello stato S . L'accuratezza della comprensione è proporzionale alla fedeltà della simulazione — maggiore per azioni/emozioni che il soggetto è in grado di eseguire. → *Sezione IV.3, Appendice H.2*

Proposizione XXIV.3 (Irriducibilità del divario esplicativo). Nessuna descrizione fisica del sistema neurobiologico — per quanto completa — implica logicamente la descrizione fenomenologica del corpo vissuto. La distanza tra le due descrizioni è strutturale: è la versione biologica del teorema di incompletezza di Gödel applicato alla conoscenza in prima persona. → *Sezione VIII*

Appendice M — Programmi di ricerca aperti

M.1 Congetture formali del terzo movimento

Congettura M.1 (Convergenza empirica della IIT). La struttura concettuale CES del sistema talamocorticale è computabile con approssimazioni scalabili (per esempio, attraverso metodi Monte Carlo per la stima di Φ in sistemi di dimensionalità alta) e la convergenza empirica della misura PCI al valore teorico di Φ/Φ_{max} è dimostrabile. *Stato: aperta; dipende da progressi computazionali nella stima di EMD in alta dimensionalità.*

Congettura M.2 (Completezza della embodied simulation per le emozioni di base). Per le sei emozioni di base (Ekman, 1972: gioia, paura, disgusto, tristezza, rabbia, sorpresa), la simulazione incarnata mediata dal sistema specchio è sufficiente per la loro comprensione — nel senso che la connessione tra riconoscimento dell'espressione altrui e attivazione del proprio sistema emotivo ha la struttura dell'inferenza bayesiana con likelihood calcolato dalla simulazione motoria/viscerale. *Stato: parzialmente verificata per disgusto (Wicker et al. 2003) e dolore (Singer et al. 2004); non ancora estesa a tutte le emozioni di base.*

Congettura M.3 (Continuità del paesaggio affettivo). La funzione di potenziale affettivo $U(v,a)$ è differenziabile in tutto il dominio per individui senza disturbi dell'umore, e ha punti di non differenziabilità localizzati attorno agli eventi emotivi intensi per individui con disturbi dell'umore. *Stato: verificabile con dati EMA ad alta risoluzione temporale.*

M.2 Connessioni non esplorate tra i due cicli

Il presente ciclo ha costruito la neurobiologia dell'intelligenza biologica come complemento formale alla teoria computazionale sviluppata nel ciclo precedente. Tre connessioni tra i due cicli rimangono parzialmente non sviluppate:

La prima è la connessione tra il principio di energia libera variazionale del cervello bayesiano (Monografia XVII del presente ciclo) e il principio di termodinamica dell'informazione (Monografia XXI del ciclo precedente). Entrambi minimizzano grandezze che si chiamano «energia libera» ma derivate da tradizioni diverse (la termodinamica statistica e l'inferenza variazionale). La relazione formale precisa tra le due grandezze — il cervello come sistema fisico che minimizza l'entropia termodinamica mentre minimizza l'energia libera variazionale — non è stata sviluppata.

La seconda è la connessione tra la teoria funzionale della decisione (FDT, Monografia XVII del ciclo precedente) e la neurobiologia della valutazione nei gangli della base (Monografie IX e XXI del presente ciclo). Il TD-learning striatale implementa qualcosa di funzionalmente analogo alla FDT — la selezione dell'algoritmo invece dell'azione individuale — ma la connessione formale precisa tra il calcolo $do(Alg = \pi)$ della FDT e il circuito actor-critic con eligibility traces non è stata formalizzata.

La terza è la connessione tra la semantica evolutiva (Monografia XIX del presente ciclo) e il problema del grounding semantico (Monografia XX del ciclo precedente). La catena causale di trasmissione kripkiana ha un precursore biologico nella trasmissione epigenetica, ma il passaggio formale preciso dalla struttura epigenetica alla struttura linguistica della

catena kripkiana — il modo in cui l'evoluzione biologica ha prodotto la capacità di trasmissione intenzionale del riferimento — non è stato derivato.

Bibliografia del ciclo *Ai confini del vivente*

La bibliografia è organizzata in ordine alfabetico per cognome dell'autore principale. Le abbreviazioni: J Neurosci = Journal of Neuroscience; PNAS = Proceedings of the National Academy of Sciences; Nature Neurosci = Nature Neuroscience; Neuron = Neuron; J Physiol = Journal of Physiology; Trends Neurosci = Trends in Neurosciences; Curr Biol = Current Biology; Science = Science; Nature = Nature; Cell = Cell; Phil Trans = Philosophical Transactions of the Royal Society; Cognition = Cognition; Psychol Rev = Psychological Review; J Cogn Neurosci = Journal of Cognitive Neuroscience; PLoS Biol = PLoS Biology; eLife = eLife; Cereb Cortex = Cerebral Cortex; J Neurophysiol = Journal of Neurophysiology; Brain Struct Funct = Brain Structure and Function.

Adams, Stephan, Brown, Frith e Friston (2013) Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D. e Friston, K.J. «The Computational Anatomy of Psychosis», *Frontiers in Psychiatry* 4, 47.

Anfinsen (1973) Anfinsen, C.B. «Principles that Govern the Folding of Protein Chains», *Science* 181(4096), 223–230.

Axelrod (1984) Axelrod, R. *The Evolution of Cooperation*. Basic Books.

Barwise e Perry (1983) Barwise, J. e Perry, J. *Situations and Attitudes*. MIT Press.

Bastos, Usrey, Adams, Mangun, Fries e Friston (2012) Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P. e Friston, K.J. «Canonical Microcircuits for Predictive Coding», *Neuron* 76(4), 695–711.

Bastos, Vezoli, Bosman, Schoffelen, Oostenveld, Dowdall, De Weerd, Kennedy e Fries (2015) Bastos, A.M. et al. «Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels», *Neuron* 85(2), 390–401.

Berkes, Orbán, Lengyel e Fiser (2011) Berkes, P., Orbán, G., Lengyel, M. e Fiser, J. «Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment», *Science* 331(6013), 83–87.

Block (1995) Block, N. «On a Confusion about a Function of Consciousness», *Behavioral and Brain Sciences* 18(2), 227–287.

Block (2007) Block, N. «Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience», *Behavioral and Brain Sciences* 30(5–6), 481–548.

Bodnar, Kim, Orr, Bhatt, Adams, Kim, Martin e Wright (1998) Bodnar, A.G., Ouellette, M., Frolkis, M., Holt, S.E., Chiu, C.P., Morin, G.B., Harley, C.B., Shay, J.W., Lichtsteiner, S. e Wright, W.E. «Extension of Life-Span by Introduction of Telomerase into Normal Human Cells», *Science* 279(5349), 349–352.

Brunet, Orr, Tremblay, Robertson, Nader e Pitman (2008) Brunet, A., Orr, S.P., Tremblay, J., Robertson, K., Nader, K. e Pitman, R.K. «Effect of Post-Retrieval Propranolol on Psychophysiologic Responding during Subsequent Script-Driven Traumatic Imagery in Post-Traumatic Stress Disorder», *Journal of Psychiatric Research* 42(6), 503–506.

Casali, Gosseries, Rosanova, Boly, Sarasso, Casali, Casarotto, Bruno, Laureys, Tononi e Massimini (2013) Casali, A.G. et al. «A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior», *Science Translational Medicine* 5(198), 198ra105.

Casarotto, Comanducci, Rosanova, Sarasso, Fecchio, Napolitani, Pigorini, Gosseries, Bodart, Tassou, Boly, Laureys, Tononi e Massimini (2016) Casarotto, S. et al. «Stratification of Unresponsive Patients by an Independently Validated Index of Brain Complexity», *Annals of Neurology* 80(5), 718–729.

Chalmers (1995) Chalmers, D.J. «Facing Up to the Problem of Consciousness», *Journal of Consciousness Studies* 2(3), 200–219.

Chalmers (1996) Chalmers, D.J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Cheney e Seyfarth (1980) Cheney, D.L. e Seyfarth, R.M. «Vocal Recognition in Free-Ranging Vervet Monkeys», *Animal Behaviour* 28(2), 362–367.

Cresswell (1985) Cresswell, M.J. *Structured Meanings: The Semantics of Propositional Attitudes*. MIT Press.

Damasio (1994) Damasio, A.R. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.

Dapretto, Davies, Pfeifer, Scott, Sigman, Bookheimer e Iacoboni (2006) Dapretto, M., Davies, M.S., Pfeifer, J.H., Scott, A.A., Sigman, M.,

Bookheimer, S.Y. e Iacoboni, M. «Understanding Emotions in Others: Mirror Neuron Dysfunction in Children with Autism Spectrum Disorders», *Nature Neuroscience* 9(1), 28–30.

Davidson (1967) Davidson, D. «The Logical Form of Action Sentences», in N. Rescher (ed.), *The Logic of Decision and Action*, University of Pittsburgh Press, 81–95.

Davidson (1970) Davidson, D. «Mental Events», in L. Foster e J.W. Swanson (eds.), *Experience and Theory*, Duckworth, 79–101.

Davidson (1980) Davidson, D. *Essays on Actions and Events*. Oxford University Press.

Dehaene e Changeux (2011) Dehaene, S. e Changeux, J.P. «Experimental and Theoretical Approaches to Conscious Processing», *Neuron* 70(2), 200–227.

Di Paolo, Rohde e De Jaegher (2010) Di Paolo, E., Rohde, M. e De Jaegher, H. «Horizons for the Enactive Mind: Values, Social Interaction, and Play», in J. Stewart, O. Gapenne e E. Di Paolo (eds.), *Enaction: Toward a New Paradigm for Cognitive Science*, MIT Press, 33–87.

Dunbar (1992) Dunbar, R.I.M. «Neocortex Size as a Constraint on Group Size in Primates», *Journal of Human Evolution* 22(6), 469–493.

Dunbar (1998) Dunbar, R.I.M. «The Social Brain Hypothesis», *Evolutionary Anthropology* 6(5), 178–190.

Dunbar, Marriott e Duncan (1997) Dunbar, R.I.M., Marriott, A. e Duncan, N.D.C. «Human Conversational Behavior», *Human Nature* 8(3), 231–246.

Ekman (1992) Ekman, P. «An Argument for Basic Emotions», *Cognition and Emotion* 6(3–4), 169–200.

Elowitz, Levine, Siggia e Swain (2002) Elowitz, M.B., Levine, A.J., Siggia, E.D. e Swain, P.S. «Stochastic Gene Expression in a Single Cell», *Science* 297(5584), 1183–1186.

Enard, Gehre, Hammerschmidt, Hölter, Blass, Somel, Brückner, Schreiweis, Winter, Sohr, Becker, Wiebe, Nickel, Giger, Müller, Groszer, Adler, Aguilar, Walter, Grüneberg, Bhatt, Bhatt, Bhatt, Bhatt, Bhatt e Pääbo (2009) Enard, W. et al. «A Humanized Version of Foxp2 Affects Cortico-Basal Ganglia Circuits in Mice», *Cell* 137(5), 961–971.

Ernst e Banks (2002) Ernst, M.O. e Banks, M.S. «Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion», *Nature* 415(6870), 429–433.

Feldman e Friston (2010) Feldman, H. e Friston, K.J. «Attention, Uncertainty, and Free-Energy», *Frontiers in Human Neuroscience* 4, 215.

Fitch e Hauser (2004) Fitch, W.T. e Hauser, M.D. «Computational Constraints on Syntactic Processing in a Nonhuman Primate», *Science* 303(5656), 377–380.

Frey e Morris (1997) Frey, U. e Morris, R.G.M. «Synaptic Tagging and Long-Term Potentiation», *Nature* 385(6616), 533–536.

Friston (2005) Friston, K. «A Theory of Cortical Responses», *Philosophical Transactions of the Royal Society B* 360(1456), 815–836.

Friston (2010) Friston, K. «The Free-Energy Principle: A Unified Brain Theory?», *Nature Reviews Neuroscience* 11(2), 127–138.

Gallese (2001) Gallese, V. «The "Shared Manifold" Hypothesis: From Mirror Neurons to Empathy», *Journal of Consciousness Studies* 8(5–7), 33–50.

Gallese e Goldman (1998) Gallese, V. e Goldman, A. «Mirror Neurons and the Simulation Theory of Mind-Reading», *Trends in Cognitive Sciences* 2(12), 493–501.

Gallagher (2007) Gallagher, S. «Simulation Trouble», *Social Neuroscience* 2(3–4), 353–365.

Greene, Sommerville, Nystrom, Darley e Cohen (2001) Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M. e Cohen, J.D. «An fMRI Investigation of Emotional Engagement in Moral Judgment», *Science* 293(5537), 2105–2108.

Grice (1957) Grice, H.P. «Meaning», *Philosophical Review* 66(3), 377–388.

Harley, Futcher e Greider (1990) Harley, C.B., Futcher, A.B. e Greider, C.W. «Telomeres Shorten during Ageing of Human Fibroblasts», *Nature* 345(6274), 458–460.

Hauser, Chomsky e Fitch (2002) Hauser, M.D., Chomsky, N. e Fitch, W.T. «The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?», *Science* 298(5598), 1569–1579.

Hayflick (1961) Hayflick, L. e Moorhead, P.S. «The Serial Cultivation of Human Diploid Cell Strains», *Experimental Cell Research* 25(3), 585–621.

Hayflick (1965) Hayflick, L. «The Limited in Vitro Lifetime of Human Diploid Cell Strains», *Experimental Cell Research* 37(3), 614–636.

Hebb (1949) Hebb, D.O. *The Organization of Behavior: A Neuropsychological Theory*. Wiley.

Helfrich, Lendner, Mander, Romundstad, Deuker, Tackle, Frank, Walker e Knight (2018) Helfrich, R.F., Lendner, J.D., Mander, B.A., Romundstad, L., Deuker, L., Tackle, A., Frank, M.J., Walker, M.P. e Knight, R.T. «Bidirectional Prefrontal-Hippocampal Dynamics Organize Information Transfer during Sleep for Memory Consolidation», *Nature Communications* 9, 5116.

Hickok e Poeppel (2004) Hickok, G. e Poeppel, D. «Dorsal and Ventral Streams: A Framework for Understanding Aspects of the Functional Anatomy of Language», *Cognition* 92(1–2), 67–99.

Huys, Daw e Dayan (2015) Huys, Q.J.M., Daw, N.D. e Dayan, P. «Depression: A Decision-Theoretic Analysis», *Annual Review of Neuroscience* 38, 1–23.

Insel (1997) Insel, T.R. «A Neurobiological Basis of Social Attachment», *American Journal of Psychiatry* 154(6), 726–735.

Johansson (1973) Johansson, G. «Visual Perception of Biological Motion and a Model for Its Analysis», *Perception and Psychophysics* 14(2), 201–211.

Kahneman e Tversky (1973) Kahneman, D. e Tversky, A. «On the Psychology of Prediction», *Psychological Review* 80(4), 237–251.

Kapur (2003) Kapur, S. «Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia», *American Journal of Psychiatry* 160(1), 13–23.

Kok, Jehee e de Lange (2012) Kok, P., Jehee, J.F.M. e de Lange, F.P. «Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex», *Neuron* 75(2), 265–270.

Kripke (1972) Kripke, S.A. *Naming and Necessity*. Harvard University Press.

Kuppens, Allen e Sheeber (2010) Kuppens, P., Allen, N.B. e Sheeber, L.B. «Emotional Inertia and Psychological Maladjustment», *Psychological Science* 21(7), 984–991.

Lai, Fisher, Hurst, Vargha-Khadem e Monaco (2001) Lai, C.S.L., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F. e Monaco, A.P. «A Forkhead-Domain Gene Is Mutated in a Severe Speech and Language Disorder», *Nature* 413(6855), 519–523.

Lamme (2006) Lamme, V.A.F. «Towards a True Neural Stance on Consciousness», *Trends in Cognitive Sciences* 10(11), 494–501.

LeDoux (1993) LeDoux, J.E. «Emotional Memory Systems in the Brain», *Behavioural Brain Research* 58(1–2), 69–79.

LeDoux (2015) LeDoux, J.E. *Anxious: Using the Brain to Understand and Treat Fear and Anxiety*. Viking.

Lee e Wilson (2002) Lee, A.K. e Wilson, M.A. «Memory of Sequential Experience in the Hippocampus during Slow Wave Sleep», *Neuron* 36(6), 1183–1194.

Levine (1983) Levine, J. «Materialism and Qualia: The Explanatory Gap», *Pacific Philosophical Quarterly* 64(4), 354–361.

Lewis, Rezaie, Brown, Roberts e Dunbar (2011) Lewis, P.A., Rezaie, R., Brown, R., Roberts, N. e Dunbar, R.I.M. «Ventromedial Prefrontal Volume Predicts Understanding of Others and Social Network Size», *NeuroImage* 57(4), 1624–1629.

Lindahl (1993) Lindahl, T. «Instability and Decay of the Primary Structure of DNA», *Nature* 362(6422), 709–715.

Loftus (1979) Loftus, E.F. e Palmer, J.C. «Reconstruction of Automobile Destruction: An Example of the Interaction between Language and Memory», *Journal of Verbal Learning and Verbal Behavior* 13(5), 585–589.

López-Otín, Blasco, Partridge, Serrano e Kroemer (2013) López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M. e Kroemer, G. «The Hallmarks of Aging», *Cell* 153(6), 1194–1217.

Lycan (1996) Lycan, W.G. *Consciousness and Experience*. MIT Press.

Markov, Ercsey-Ravasz, Van Essen, Knoblauch, Toroczkai e Kennedy (2014) Markov, N.T. et al. «A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex», *Cerebral Cortex* 24(1), 17–36.

Marr (1982) Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman.

McCloskey e Cohen (1989) McCloskey, M. e Cohen, N.J. «Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem», *Psychology of Learning and Motivation* 24, 109–165.

Mehta, Lee e Wilson (2002) Mehta, M.R., Lee, A.K. e Wilson, M.A. «Role of Experience and Oscillations in Transforming a Rate Code into a Temporal Code», *Nature* 417(6890), 741–746.

Melloni, Mudrik, Pitts, Bendtz, Ferrante, Gorman, Hirschhorn, Khalaf, Lepauvre, Lichtner, Michelmann, Miyamoto, Moyal, Mudrik, Northoff, Ouyang, Slater, Soto, Sterzer, Sugahara, Timmermann, Tsuchiya e Koch (2023) Melloni, L. et al. «An Adversarial Collaboration Protocol for Empirically Adjudicating between Theories of Consciousness», *eLife* 12, e81242.

Merleau-Ponty (1945) Merleau-Ponty, M. *Phénoménologie de la perception*. Gallimard. Trad. it.: *Fenomenologia della percezione*, Il Saggiatore, 2003.

Merleau-Ponty (1968) Merleau-Ponty, M. *Le Visible et l'Invisible*. Gallimard. Trad. it.: *Il visibile e l'invisibile*, Bompiani, 1969.

Meyerhans, Cheynier, Albert, Seth, Kwok, Evans, Desrosiers, Bhatt, Bhatt e Bhatt (1989) Meyerhans, A., Cheynier, R., Albert, J., Seth, M., Kwok, S., Bhatt, R., Bhatt, J.N., Bhatt, D.L., Norrby, E. e Wain-Hobson, S. «Temporal Fluctuations in HIV Quasispecies in Vivo Are Not Reflected by Sequential HIV Isolations», *Cell* 58(5), 901–910.

Montague (1970) Montague, R. «Universal Grammar», *Theoria* 36(3), 373–398.

Montague (1973) Montague, R. «The Proper Treatment of Quantification in Ordinary English», in J. Hintikka, J. Moravcsik e P. Suppes (eds.), *Approaches to Natural Language*, Reidel, 221–242.

Moser, Kropff e Moser (2008) Moser, E.I., Kropff, E. e Moser, M.B. «Place Cells, Grid Cells, and the Brain's Spatial Representation System», *Annual Review of Neuroscience* 31, 69–89.

Nader, Schafe e LeDoux (2000) Nader, K., Schafe, G.E. e LeDoux, J.E. «Fear Memories Require Protein Synthesis in the Amygdala for Reconsolidation after Retrieval», *Nature* 406(6797), 722–726.

Nadel e Moscovitch (1997) Nadel, L. e Moscovitch, M. «Memory Consolidation, Retrograde Amnesia and the Hippocampal Complex», *Current Opinion in Neurobiology* 7(2), 217–227.

Nádasdy, Hirase, Czurkó, Csicsvari e Buzsáki (1999) Nádasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J. e Buzsáki, G. «Replay and Time Compression of Recurring Spike Sequences in the Hippocampus», *Journal of Neuroscience* 19(21), 9497–9507.

Nagel (1974) Nagel, T. «What Is It Like to Be a Bat?», *Philosophical Review* 83(4), 435–450.

Niedenthal, Brauer, Halberstadt e Innes-Ker (2001) Niedenthal, P.M., Brauer, M., Halberstadt, J.B. e Innes-Ker, Å.H. «When Did Her Smile Drop? Facial Mimicry and the Influences of Emotional State on the Detection of Change in Emotional Expression», *Cognition and Emotion* 15(6), 853–864.

Nowak e May (2000) Nowak, M.A. e May, R.M. *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press.

Oberman, Hubbard, McCleery, Altschuler, Ramachandran e Pineda (2005) Oberman, L.M. et al. «EEG Evidence for Mirror Neuron Dysfunction in Autism Spectrum Disorders», *Cognitive Brain Research* 24(2), 190–198.

O'Keefe e Dostrovsky (1971) O'Keefe, J. e Dostrovsky, J. «The Hippocampus as a Spatial Map: Preliminary Evidence from Unit Activity in the Freely Moving Rat», *Brain Research* 34(1), 171–175.

Oja (1982) Oja, E. «A Simplified Neuron Model as a Principal Component Analyzer», *Journal of Mathematical Biology* 15(3), 267–273.

Pearl (2000) Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Quine (1960) Quine, W.V.O. *Word and Object*. MIT Press.

Rizzolatti, Fadiga, Gallese e Fogassi (1996) Rizzolatti, G., Fadiga, L., Gallese, V. e Fogassi, L. «Premotor Cortex and the Recognition of Motor Actions», *Cognitive Brain Research* 3(2), 131–141.

Rosenthal (1997) Rosenthal, D.M. «A Theory of Consciousness», in N. Block, O. Flanagan e G. Güzeldere (eds.), *The Nature of Consciousness*, MIT Press, 729–753.

Rosenthal (2005) Rosenthal, D.M. *Consciousness and Mind*. Oxford University Press.

Scherer (2009) Scherer, K.R. «The Dynamic Architecture of Emotion: Evidence for the Component Process Model», *Cognition and Emotion* 23(7), 1307–1351.

Schultz, Dayan e Montague (1997) Schultz, W., Dayan, P. e Montague, P.R. «A Neural Substrate of Prediction and Reward», *Science* 275(5306), 1593–1599.

Seyfarth, Cheney e Marler (1980) Seyfarth, R.M., Cheney, D.L. e Marler, P. «Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication», *Science* 210(4471), 801–803.

Singer, Seymour, O'Doherty, Kaube, Dolan e Frith (2004) Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J. e Frith, C.D. «Empathy for Pain Involves the Affective but Not Sensory Components of Pain», *Science* 303(5661), 1157–1162.

Staresina, Bergmann, Bonnefond, van der Meij, Jensen, Deuker, Elger, Axmacher e Fell (2015) Staresina, B.P. et al. «Hierarchical Nesting of Slow Oscillations, Spindles and Ripples in the Human Hippocampus during Sleep», *Nature Neuroscience* 18(11), 1679–1686.

Sterling (2012) Sterling, P. «Allostasis: A Model of Predictive Regulation», *Physiology and Behavior* 106(1), 5–15.

Steyvers e Tenenbaum (2005) Steyvers, M. e Tenenbaum, J.B. «The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth», *Cognitive Science* 29(1), 41–78.

Tononi (2004) Tononi, G. «An Information Integration Theory of Consciousness», *BMC Neuroscience* 5, 42.

Tononi, Boly, Massimini e Koch (2016) Tononi, G., Boly, M., Massimini, M. e Koch, C. «Integrated Information Theory: From Consciousness to Its Physical Substrate», *Nature Reviews Neuroscience* 17(7), 450–461.

Turing (1952) Turing, A.M. «The Chemical Basis of Morphogenesis», *Philosophical Transactions of the Royal Society B* 237(641), 37–72.

van der Veen, Murphy, den Hartog, Bhatt, Bhatt, Bhatt, Bhatt e Bhatt (2012) van der Veen, F.M. et al. «Effects of Acute Tryptophan Depletion on Emotional Processing in Unmedicated and SSRIs-Treated Patients with Major Depressive Disorder», *Neuropsychopharmacology* 37(13), 2835–2842.

Varela, Thompson e Rosch (1991) Varela, F.J., Thompson, E. e Rosch, E. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Vargha-Khadem, Watkins, Alcock, Fletcher e Passingham (1995) Vargha-Khadem, F., Watkins, K., Alcock, K., Fletcher, P. e Passingham, R. «Praxic and Nonverbal Cognitive Deficits in a Large Family with a Genetically Transmitted Speech and Language Disorder», *PNAS* 92(3), 930–933.

Watson (1972) Watson, J.D. «Origin of Concatemeric T7 DNA», *Nature New Biology* 239(94), 197–201.

Weiss, Simoncelli e Adelson (2002) Weiss, Y., Simoncelli, E.P. e Adelson, E.H. «Motion Illusions as Optimal Percepts», *Nature Neuroscience* 5(6), 598–604.

Wicker, Keysers, Plailly, Royet, Gallese e Rizzolatti (2003) Wicker, B., Keysers, C., Plailly, J., Royet, J.P., Gallese, V. e Rizzolatti, G. «Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust», *Neuron* 40(3), 655–664.

Wilson e McNaughton (1994) Wilson, M.A. e McNaughton, B.L. «Reactivation of Hippocampal Ensemble Memories during Sleep», *Science* 265(5172), 676–679.

Wimmer e Perner (1983) Wimmer, H. e Perner, J. «Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception», *Cognition* 13(1), 103–128.

Wolpert e Macready (1997) Wolpert, D.H. e Macready, W.G. «No Free Lunch Theorems for Optimization», *IEEE Transactions on Evolutionary Computation* 1(1), 67–82.

Ron Fong e Greider (1984) Greider, C.W. e Blackburn, E.H. «Identification of a Specific Telomere Terminal Transferase Activity in Tetrahymena Extracts», *Cell* 43(2), 405–413.